Powerful Association Testing with Application to Neuroimaging Genetics

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Zhiyuan Xu

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

Advised by Wei Pan, Ph.D

May, 2017

Acknowledgements

First, I would like to express my sincere gratitude and appreciation to my advisor, Professor Wei Pan, for his support and guidance not only on my research, but also many other aspects of my life. I could not have imagined having a better advisor during my Ph.D study.

Besides my advisor, I would like to extend my profound gratitude to the rest of my thesis committee: Prof. Saonli Basu, Prof. Baolin Wu and Prof. Lan Wang for their precious time and constructive comments on my thesis.

Abstract

In spite of the huge success of the standard single-nucleotide polymorphism (SNP) based analysis in genome-wide association studies (GWASs), it has some limitations. First, it suffers power loss from a stringent significance level due to multiplicity adjustment for up to millions of tests. In addition, it has low power since the effect sizes of SNPs are usually small. Instead, gene-based testing might improve statistical power by aggregating moderate to weakly associated SNPs within each gene while greatly reducing the burden of multiple testing adjustment from millions to thousands. Second, almost all existing analyses do not explicitly account for (unknown) genetic heterogeneity, leading to possible loss of power as convincingly shown in simulation studies (Londono et al., 2012; Qian and Shao, 2013; Zhou and Pan, 2009). Moreover, as there are many other data resources available (e.g. neuroimaging phenotypes, molecular phenotypes like gene expression) besides GWAS/DNA sequencing data, integrating them into GWAS is expected to boost statistical power.

We first introduce a flexible framework to extend score-based testing in generalized linear models to more complex models, for example, mixed effect models. Second, we show that by accounting for genetic heterogeneity, more associated SNPs can be detected than the standard one-degree-of-freedom trend test in single SNP-based testing. Third, we propose a new adaptive *aSPC* test to detect associations between two random vectors in moderate to high dimensions; we also point out its connections to some existing association testing for multiple SNPs and multiple traits. Finally, we propose a novel gene-based association testing approach by incorporating weights derived from other data resources (e.g. from another eQTL dataset). We show the power gain of the new approach over two existing methods *PrediXcan* and *TWAS*, pointing out that both *PrediXcan* and *TWAS* are special cases of our new test.

Contents

Acknowledgements **i**

Abstract **ii**

List of Tables **vi**

List of Figures **ix**

1 Introduction **1**

1.1 Background 1

1.2 Data 4

2 Approximate score-based testing with application to multivariate trait association analysis **5**

2.1 Introduction 5

2.2 Methods 7

2.2.1 Review: Some Score-based Tests 7

2.2.2 Two Examples Models 9

2.2.3 New Method: Approximating the Score Vector 11

2.3 Results 13

2.3.1 Example 13

2.3.2 Simulations 14

2.4 Conclusions 20

3 Binomial mixture model based association testing to account for genetic heterogeneity for GWAS **25**
3.1 Introduction . . . 25
3.2 Methods . . . 26
3.2.1 The Trend Test and Related Tests . . . 27
3.2.2 Hardy-Weinberg Equilibrium (HWE) Exact test . . . 28
3.2.3 Association Testing under Genetic Heterogeneity . . . 29
3.3 Application to the WTCCC data . . . 31
3.3.1 Quality Control . . . 31
3.3.2 GWAS Results . . . 31
3.4 Conclusions . . . 33

4 Adaptive testing for association between two random vectors in moderate to high dimensions **43**
4.1 Introduction . . . 43
4.2 Methods . . . 45
4.2.1 Review: the RV test . . . 46
4.2.2 New method: an adaptive sum of powered correlation (aSPC) test 47
4.2.3 Connections with some existing tests . . . 48
4.2.4 Extensions . . . 50
4.2.5 Software . . . 54
4.3 Real data application . . . 54
4.3.1 Testing for SNP-gene expression associations . . . 54
4.4 Simulations . . . 56
4.4.1 Simulation I: linear associations . . . 56
4.4.2 Simulation II: non-linear associations . . . 57
4.5 Conclusions . . . 58

5 A powerful framework for integrating eQTL and GWAS summary data **64**
5.1 Introduction . . . 64
5.2 Methods . . . 66
5.2.1 PrediXcan and TWAS . . . 66

5.2.2 Novel reformulation and extensions 67
5.2.3 New method: aSPU 68
5.2.4 Association testing with summary statistics 69
5.2.5 Association testing with multiple sets of weights 70
5.3 Results 71
5.3.1 Application to the WTCCC data 71
5.3.2 Application to the lipid GWAS summary data 72
5.4 Conclusions 75

6 Discussion and future work **88**

7 References **91**

List of Tables

2.1 P-values of the various mvLMM-based tests in analysis of the ADNI data. 21

2.2 Simulation I: empirical Type I error and power for two SNP-phenotype pairs, rs7526034-RLatTemp (pair 1) and APOE-ϵ4-RPar (pair 2), based on fitting mvLMM. 21

2.3 Simulation I: empirical Type I error and power for two SNP-phenotype pairs, rs7526034-RLatTemp (pair 1) and APOE-ϵ4-RPar (pair 2) based on fitting marginal linear models (LMs) with or without the top 10 PCs. 22

2.4 Simulation II: empirical type I error rates and power of the approximate (approx) score- and exact score-based tests when multiple traits were correlated with a CS structure with correlation $r = 0.3$. 22

2.5 Simulation III: empirical Type I error (for OR = 1) and power (for OR > 1) of the approximate score-based tests with correlated binary traits. 23

2.6 Simulation III: empirical Type I error (for OR = 1) and power (for OR > 1) of the GEE score-based tests with correlated binary traits. 24

3.1 The genotype frequencies for case-control data of a SNP. 33

3.2 The QC summary. For each disease study group, both case and two control groups are included. The number of subjects for the two control groups (i.e. NBS+58C) were before and after QC were 3,004 and 2,938, respectively. . . . 33

3.3 The inflation factors of various tests. 33

3.4 Example SNPs which are identified to be significant by LRT-H, but not by HWE, 1 df Trend or 2 df Genotypic test. 34

3.5 Some significant risk loci identified by HWE or LRT-H, but not by the other two tests. Each risk locus is within 250 Kb of some previously identified SNPs or loci. For each risk locus, the LD-independent SNP along with the MAF in the control group and the p-values of LRT-H ($P_{LRT\text{-}H}$) and the HWE exact test (P_{HWE}) are reported. 35

3.6 Simulation: empirical Type I error and power when the number of independent columns (denote as "No. ind") is 25, 35, and 65. "RV.asy" and "RV.perm" standards for asympototic and permutation-based RV test, respectively. 36

4.1 The analysis results for the ADNI data. p and q denote the numbers of SNPs and of probes surviving the P-value cut-off based on the corresponding univariate SNP-gene expression associations. 60

4.2 Simulation I: empirical Type I error and power rates when the number of independent columns (denoted as "No. ind") is 25, 45, and 65 respectively. "RV.asy" and "RV.perm" stand for the asymptotic and permutation-based RV tests, respectively. 61

5.1 Significant genes identified by the aSPU test, but not by the SPU(1) test (or PrediXcan) at the genome-wide significance threshold of 5.56×10^{-6}. The validated gene-trait associations appeared in the following references: [1] Franke et al (2010); [2] Kenny et al (2012); [3] Plagnol et al (2011). 86

5.2 The numbers of the significant genes identified by analyzing the 2010 lipid data for each single set of the weights and the combined one (i.e. with the omnibus aSPU and TWAS tests). The numbers a/b/c in each cell indicate the numbers of (a) the significant genes; (b) the significant genes that covered a genome-wide significant SNPs in the 2010 lipid data; (c) the significant genes that covered a genome-wide significant SNPs in the 2013 lipid data. 86

5.3 The numbers of the significant genes identified by analyzing the 2013 lipid data for each single set of the weights and the combined one (i.e. with the omnibus aSPU and TWAS tests). The numbers a/b/c in each cell indicate the numbers of (a) the significant genes; (b) the significant genes that covered a genome-wide significant SNPs in the 2010 lipid data; (c) the significant genes that covered a genome-wide significant SNPs in the 2013 lipid data. 87

5.4 Significant gene-trait associations identified by aSPU or/and TWAS with no known risk loci within 500kb. The column "Sig.test" indicates the corresponding association was detected by aSPU or SPU(1) or both. . . 87

List of Figures

3.1 Linkage disequilibrium plots for simulated genotypes with ($n = 200$, $p = 10$) (left panel) and ($n = 400$, $p = 20$) (right panel). 36

3.2 Q-Q plots of various tests for CD (first row), BD (2nd row), CAD (3rd row) and T2D (bottom row). 37

3.3 Venn-diagrams of the significant SNPs at the genome-wide significance level of 5×10^{-8} identified by each test for traits CD, BD, CAD and T2D (from the top to the bottom). 38

3.4 Venn-diagrams of the significant risk loci identified by each test for traits CD, BD, CAD and T2D (from the top to the bottom). 39

3.5 LocusZoom plots of the risk loci for trait CD, uniquely identified by LRT-H. The GenoCanyon scores for the LD-independent (index) SNPs are 2.71e-05, 5.3e-04, 1.00, 2.4e-03 and 5.5e-03 respectively. 40

3.6 LocusZoom plots of the risk loci for trait BD, uniquely identified by LRT-H. The GenoCanyon scores for the LD-independent (index) SNPs are 0.999, 0.888 and 0.972 respectively. 41

3.7 LocusZoom plots of the risk loci for trait CAD, uniquely identified by LRT-H . The GenoCanyon scores for the LD-independent (index) SNPs are 3.35e-06 and 1.24e-06 respectively. 42

3.8 LocusZoom plot of the risk locus for trait T2D, uniquely identified by LRT-H. The GenoCanyon score for the LD-independent (index) SNP is 5.34e-05. . . . 42

4.1 The computing time of the permutation-based RV, GEE-aSPU, aSPC.P, aSPC.Sp and aSPC.dCor tests. The left panel shows the computing time of aSPC.dCor test as compared to that of all the other tests, while the right panel is a zoom-in for all the tests except aSPC.dCor. 62

4.2 Simulation II results. The left panel: when the number of columns in X and Y are 5, the empirical type I error and power curves of the tests as the number of truly non-linearly associated column pairs between X and Y ranges from 0 (type I error) to 5. Right panel: when the number of non-linearly associated column pairs in X and Y is fixed at 5, the power curves of the tests as more and more non-associated columns are added to Y. The nominal significance level is 0.05. 63

5.1 The Manhattan plots for the pooled results of aSPU and aSPU-O for traits HDL (top) and LDL (bottom) based on the 2013 lipid data. The letters "(n)", "(y)", "(m)" and "(o)" following a gene's name indicate the result of aSPU based on the NTR, YFS and METSIM weights and that of aSPU-O respectively. 76

5.2 The Manhattan plots for the pooled results of aSPU and aSPU-O for traits TG (top) and TC (bottom) based on the 2013 lipid data. 77

5.3 The scatter plots of the base 10 $-\log$(p-values) of the SPU(1) test and PrediXcan applied to the WTCCC data with weights derived from the DGN whole blood for traits CD, BD, CAD, RA, HT, T1D and T2D (from the left to the right and the top to the bottom). The Pearson correlation coefficient between the two sets of the -log p-values in each panel was equal to 1. For better visualization, all the p-values were truncated at 1×10^{-9}. 78

5.4 The Q-Q plots applied to the WTCCC data with the weights derived from the DGN whole blood gene expression for traits CD, BD, CAD, RA, HT, T1D and T2D (from the left to the right and the top to the bottom). The second column in each legend indicates the numbers of the significant genes identified at the genome-wide significance level of 5.56×10^{-6}. For better visualization, the p-values of aSPU were truncated at 1×10^{-7}, while those of the other two asympototic tests were truncated at 1×10^{-9}. 79

5.5 The scatter plots of the base 10 $-\log$(p-values) of the SPU(1) test and TWAS applied to 2010 lipid data with each set of the weights based on the NTR, YFS and METSIM studies, corresponding to the 1st, 2nd and 3rd columns, respectively. The panels from the top row to the bottom row correspond to traits HDL, LDL, TC and TG, respectively. The Pearson correlation coefficients in each panel was equal to 1. For better visualization, all the p-values were truncated at 1×10^{-9}. 80

5.6 The Q-Q plots for the 2010 lipid data with each of the weights NTR, YFS and METSIM, corresponding to the 1st, 2nd and 3rd columns, respectively. The panels from the top row to the bottom row correspond to HDL, LDL, TC and TG, respectively. For better visualization, the p-values of aSPU were truncated at 1×10^{-7}, while those of the other two asympototic tests were truncated at 1×10^{-9}. 81

5.7 The Q-Q plots for the 2013 lipid data with each of the weights NTR, YFS and METSIM, corresponding to the 1st, 2nd and 3rd columns, respectively. The panels from the top row to the bottom row correspond to HDL, LDL, TC and TG, respectively. For better visualization, the p-values of aSPU were truncated at 1×10^{-7}, while those of the other two asympototic tests were truncated at 1×10^{-9}. 82

5.8 The Manhattan plots of aSPU (top) and TWAS (bottom) applied to the 2013 lipid dataset with trait LDL and the YFS-based weights. The p-values of both aSPU and TWAS were truncated at 1×10^{-7}. 83

5.9 The Manhattan plots of aSPU-O (top) and TWAS-O (bottom) applied to the 2013 lipid dataset for trait LDL and combining over the three sets of the weights. The p-values of both aSPU-O and TWAS-O were truncated at 1×10^{-7}. 84

5.10 The Q-Q plots for the WTCCC genotypic data and a randomly simulated binary trait. (a) Results with the individual level data and the weights derived from the DGN whole blood gene expression; (b-d) the results with the summary statistics and the weights derived from the NTR, YFS and METSIM studies, respectively; (e) the results with the summary statistics and combining the three sets of the weights. For (a) - (d), the three numbers in each parenthesis correspond to inflation factors of aSPU, SPU(1) and SPU(2), respectively; for (e), the number in the parenthesis corresponds to the inflation factor of aSPU-O. For better visualization, the p-values of aSPU were truncated at 1×10^{-7}, while those of the other two asympototic tests were truncated at 1×10^{-9}. 85

Chapter 1

Introduction

1.1 Background

Neuroimaging genetics has been increasingly drawn attention because of its usefulness to understand the mechanism of diseases. For example, Alzhemer's Disease Neuroimaging Initiative (ADNI) is a national-wide longitudinal study with large amount of clinical, imaging, genetic and biochemical biomarker data collected for the early detection and tracking of Alzheimer's disease (AD). This thesis focuses on developing powerful statistical testing methods to detect associations between genetic variants and phenotypes of interest (e.g. disease status or neuroimaging traits).

In spite of the success of single SNP association analysis, it might be too conservative as there are hundreds of millions tests needed to be adjusted. Instead, gene-based testing might improve power as it greatly reduces the burden of multiple tests adjustment from millions to thousands. In neuroimaging studies, hundreds and thousands of secondary neuroimaging phenotypes are measured (e.g. grey matter density of a region of interest in human brain); conducting association analysis for these multiple and correlated phenotypes are of interest.

We first point out the limitation of existing score-based testings and extend them to complex models by approximating score vectors. For genome-wide association studies and DNA sequencing studies, several powerful score-based tests, such as kernel machine regression and sum of powered score tests, have been proposed in the last few years.

However, extensions of these score-based tests to more complex models, such as mixed-effects models for analysis of multiple and correlated traits, have been hindered by the unavailability of the score vector, due to either no output from statistical software or no closed-form solution at all. We propose a simple and general method to asymptotically approximate the score vector based on an asymptotically normal and consistent estimate of a parameter vector to be tested and its (consistent) covariance matrix. The proposed method is applicable to both maximum-likelihood estimation and estimating function-based approaches. We use the derived approximate score vector to extend several score-based tests to mixed-effects models. We demonstrate the feasibility and possible power gains of these tests in association analysis of multiple and correlated quantitative or binary traits with both real and simulated data. The proposed method is easy to implement with a wide applicability.

Second we emphasize the potential usefulness of accounting for genetic heterogeneity in GWASs. GWASs have confirmed the ubiquitous existence of genetic heterogeneity for common disease: multiple common genetic variants have been identified to be associated, while many more are yet expected to be uncovered. On the other hand, the single SNP-based trend test (or its variants) that has been dominantly used in GWASs is based on contrasting the allele frequency difference between the case and control groups, completely ignoring possible genetic heterogeneity. In spite of the widely accepted notion of genetic heterogeneity, we are not aware of any previous attempt to apply genetic heterogeneity-motivated methods in GWAS. Here, to explicitly account for unknown genetic heterogeneity, we applied a mixture model-based single SNP test to the WTCCC GWAS data with traits Crohn's disease, bipolar disease, coronary artery disease and type 2 diabetes, identifying much larger numbers of significant SNPs and risk loci for each trait than those of the popular trend test, demonstrating potential power gain of the mixture model-based test.

Third we propose an association testing method between two random vectors, which can be applied in testing associations between multiple neuroimaging phenotypes and genetic variants. Testing association between two random vectors is a common and important task in many fields, however, existing tests, such as the RV test, are suitable only for low-dimensional data, not for high-dimensional data. In moderate to high dimensions, it is necessary to consider sparse signals, which are often expected with

only a few, but not many, variables associated with each other. We generalize the RV test to moderate to high dimensions. The key idea is to data-adaptively weight each variable pair based on its empirical association. As the consequence, the proposed test is adaptive, alleviating the effects of noise accumulation in high-dimensional data, and thus maintaining the power for both dense and sparse alternative hypotheses. We show the connections between the proposed test with several existing tests, such as a GEE-based adaptive test, multivariate kernel machine regression and kernel distance methods. Furthermore, we modify the proposed adaptive test so that it can be powerful for non-linear or non-monotonic associations. We use both real data and simulated data to demonstrate the advantages and usefulness of the proposed new test.

Finally we introduce a novel gene-based testing by incorporating weights from other data sources (e.g. gene-expression). Two new gene-based association analysis methods, called *PrediXcan* and *TWAS* for GWAS individual-level and summary data respectively, were recently proposed to integrate GWAS with eQTL data, alleviating two common problems in GWAS by boosting statistical power and facilitating biological interpretation of GWAS discoveries. Based on a novel reformulation of PrediXcan and TWAS, we propose a more powerful gene-based association test to integrate single set or multiple sets of eQTL data with GWAS individual-level data or summary statistics. The proposed test was applied to several GWAS datasets, including two lipid summary association datasets based on $\sim 100,000$ and $\sim 189,000$ samples respectively, and uncovered more known or novel trait-associated genes, showcasing much improved performance of our proposed method.

The rest of the thesis is organized as follows. Chapter 2 discusses the limitation of the existing score-based tests and how we can extend them to more complex models. In Chapter 3, we point out the potential power gains by accounting genetic heterogeneity in GWASs. Chapter 4 introduces a method to test association between two random vectors in moderate to high dimension, which can be applied to test associations between multi-SNP and multi-phenotypes. Finally in Chapter 5, we propose a novel gene-based association testing by incorporating weights derived from other data resources (e.g. gene-expression).

1.2 Data

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`adni.loni.usc.edu`). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California-San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date these three protocols have recruited over 1500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow up duration of each group is specified in the protocols for ADNI-1, ADNI-2 and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see `www.adni-info.org`.

Chapter 2

Approximate score-based testing with application to multivariate trait association analysis

2.1 Introduction

The main part of this chapter has been published in Xu and Pan [2015].

To detect genetic associations in genome-wide association studies (GWASs) and DNA sequencing studies, in addition to the popular univariate minimum P-value (UminP) test, many multivariate methods have been proposed to improve statistical power. Several competitive ones are score based, such as the classic score test, a variance-component score test in kernel machine regression (KMR) [Kwee et al., 2008; Wu et al., 2010, 2011], an adaptive score test [Lin and Tang, 2011], and an adaptive sum of powered score (aSPU) test [Pan et al., 2014]. A challenge is how to extend these score-based tests to more complex models beyond the generalized linear models (GLMs) for independent data. There are several reasons to consider more complex mixed-effects models in genetic association studies. First, even for a single-trait analysis, to properly account for some complex and hidden relatedness among the study subjects, or more generally for population structure or population stratification, mixed-effects models have been proposed as a general and effective approach [e.g., Yu et al., 2011; Zhang et al., 2010;

Zhou and Stephens, 2014]. These mixed-effects models differ from the standard ones in that a random effect is introduced to induce correlations among *all* the subjects, thus requiring some special and fast algorithms for model fitting as implemented in several recent software packages. These packages do not directly output the score vector. Second, there has been increasing interest in association analysis of multiple traits, which may help gain power and shed light on pleiotropy. In addition, one may encounter correlated traits as arising from familial studies. To account for correlations among multiple traits, either marginal models (based on generalized estimating equations, GEE) [Liang and Zeger, 1986] or mixed-effects models [Breslow and Clayton, 1993] can be applied. For quantitative traits, a linear mixed-effects model (LMM) can be used, from which the score vector can be derived. Accordingly the KMR test has been extended to LMMs [Maity et al., 2012; Schifano et al., 2012]. However, it is unclear how to extend the KMR and other score-based tests to GLM models (GLMMs) and Cox mixed-effects models, for which there is no closed form for the score vector (because the marginal likelihood involves an integral with random effects and in general has no closed form) [Breslow and Clayton, 1993]. Although as an alternative to GLMM, marginal models/GEE can be used, from which (generalized) score-based tests can be derived [Wang et al., 2013; Zhang et al., 2014], there may be substantial differences between the two in terms of modeling assumptions, interpretation, and thus their choices [Diggle et al., 2013]. More importantly, in genetic association studies, as discussed earlier, random effects may be necessary to effectively account for population structure, prompting the use of non-LMMs. In these situations, due to the lack of computer output or closed-form solution for the score vector, it is challenging to implement score-based tests.

In this chapter, we propose a simple and yet general method to approximate the score vector for any model. It is based on the asymptotics of an estimator of the parameters to be tested. It applies to both the maximum-likelihood estimates (MLEs) and estimating function-based estimates. Its implementation involves only a few lines of R code. We demonstrate its use in two types of mixed-effects models, a multivariate LMM (mvLMM) proposed very recently for genetic association analysis of multiple quantitative traits while correcting for cryptic relatedness and population stratification [Zhou and Stephens, 2014], and a GLMM with correlated binary traits. Among others, we use both real and simulated data to illustrate possible power gains of some approximate

score-based tests over the standard Wald and UminP tests.

2.2 Methods

2.2.1 Review: Some Score-based Tests

Suppose $U = (U_1, \ldots, U_k)'$ is the score vector for a set of k parameters to be tested with $H_0 : \psi = 0$. The classic score test is

$$T_{sco} = U^T \widehat{\mathrm{Cov}}(U)^{-1} U \tag{2.1}$$

which is aymptotically equivalent to the Wald test and likehood ratio test (LRT). The UminP test that has been widely used in GWASs can be writted as

$$T_{\mathrm{UminP}} = \max_{j=1}^{k} U_j^2 / V_{jj} \tag{2.2}$$

where V_{jj} is the jth diagonal element of $V = \widehat{\mathrm{Cov}}(U)$. Recently, a variance-component score test in KMR has been proposed for GLMs and shown to be powerful for analysis of single nucleotide polymorphism (SNP) sets [Kwee et al., 2008; Wu et al., 2010, 2011]. As discussed by Pan (2011), with a linear kernel it is equivalent to the sum of squared score (SSU) test:

$$T_{\mathrm{SSU}} = U^T U = \sum_{j=1}^{k} U_j^2. \tag{2.3}$$

Pan et al. [2014] proposed a family of the so-called SPU tests:

$$T_{\mathrm{SPU}(\gamma)} = \sum_{j=1}^{p} U_j^{\gamma} \tag{2.4}$$

for a set of integers $\gamma \geq 1$. It is easy to see that SPU(1) and SPU(2) are exactly the same as the Sum test and the SSU test, respectively [Pan, 2009]; the sum test, an example of so-called burden tests, has been shown to perform well for genetic association testing, especially for rare variants, under some situations [Li and Leal, 2008; Pan, 2009]. In

addition, for an even integer $\gamma \to \infty$, we have

$$T_{\mathrm{SPU}(\gamma)} \propto \left(\sum_{j=1}^{p} |U_j|^{\gamma} \right)^{1/\gamma} \to \max_j |U_j| = T_{\mathrm{SPU}(\infty)}, \tag{2.5}$$

the SPU(∞) is closely related to the UminP test (but ignoring possibly varying variances of U_j's). Alternatively, an SPU(γ) test can be regarded as a weighted score test [Lin and Tang, 2011] with adaptive weights $U_j^{\gamma-1}$ on each component j. In practice, because it is unknown which γ value would yield high power, we use an adaptive SPU (aSPU) test to combine the evidence across the SPU tests:

$$T_{\mathrm{aSPU}} = \min_{\gamma \in \Gamma} P_{\mathrm{SPU}(\gamma)} \tag{2.6}$$

where $P_{\mathrm{SPU}(\gamma)}$ is the P-value of the SPU(γ) test, and Γ contains a set of candidate values γ. Pan et al. [2014] found that in many situations $\Gamma = 1, 2, 3, \ldots, 8, \infty$ appeared to perform well, which will be used here.

In general, resampling methods can be used to obtain P-values for the SPU and aSPU tests. In this chapter, we assume that the asymptotic null distribution of the score vector $U \sim N(0, V)$ holds (under H_0). Accordingly, we can generate B-independent copies of the null score vector $U^{(b)}$, for which the B copies of the SPU test statistics can be calculated. Then the P-value of each SPU(γ) is calculated as $P_{\mathrm{SPU}(\gamma)} = \sum_{b=1}^{B} I(|T^{(b)}_{\mathrm{SPU}(\gamma)}| \geq |T_{\mathrm{SPU}(\gamma)}|)/B$. Furthermore, based on the same B copies of the simuated score vectors, we calculate the P-value for the aSPU test as $P_{\mathrm{aSPU}} = \sum_{b=1}^{B} I(T^{(b)}_{\mathrm{aSPU}} \leq T_{\mathrm{aSPU}})/B$ with $T^{(b)}_{\mathrm{aSPU}} = \min_{\gamma \in \Gamma} p^{(b)}_{\gamma}$ and $p^{(b_1)}_{\gamma} = \sum_{b \neq b_1} I(|T^{(b)}_{\mathrm{SPU}(\gamma)}| \geq |T^{(b_1)}_{\mathrm{SPU}(\gamma)}|)/(B-1)$.

In this chapter, we use the SPU and aSPU tests as examples, though other score-based tests [e.g., Lin and Tang, 2011; Sun et al., 2013; Wu et al., 2010] can be equally applied. Our main point does not depend on the choice of a specific score-based test; rather, we aim to show how to extend a score-based test to cases where there is no easy access to the score vector, as arising in below two important applications. To be concrete, we focus on detecting genetic association with SNPs, but the proposed method is generally applicable to other problems of interest.

2.2.2 Two Examples Models

First we need some notations, which will be used in the rest of the article unless specified otherwise.

Suppose we observe for each individual $i = 1, \ldots, n$, the response vector $Y_i = (Y_{i1}, \ldots, Y_{ik})^T$ consists of k traits (Y_i is a scalar if $k = 1$), or in familial data, we observe for each family $i = 1, \ldots, n$, the vector $Y_i = (Y_{i1}, \ldots, Y_{ik})^T$ is the response for the k members in ith family; Y_{ij} $(j = 1, \ldots, k)$ can be either quantitative or binary; $W_i = (W_{i1}, \ldots, W_{iq})^T$ is a vector including q covariates; and genetic scores for a set of p SNPs $X_i = (X_{i1}, \ldots, X_{ip})$, where $X_{ij} \in \{0, 1, 2\}, \forall j = 1, \ldots, p$; X is a $n \times p$ design matrix with ith row corresponding to X_i^T and W is a $n \times q$ design matrix with ith row of W corresponds to W_i^T. We emphasize that X_i and W_i can be either common or different across phenotypes $Y_{i1}, \ldots, Y_{ik}$, for example, in familial data, we assume X_i and W_i to be different for different family members, while in the case where we observe each individual has multiple traits, we often assume X_i and W_i to be the same for an individual. In the remainder of the manuscript, unless specified otherwise, we assume the case where we observe each individual has multiple traits and let the covariates W_i and X_i be common to phenotypes.

Multivariate linear mixed model

A multivariate linear mixed model (mvLMM) was proposed by Zhou and Stephens [2014] to test for association with multiple phenotypes while correcting for possible population stratification. Specifically, suppose we would like to test for association between a multivariate trait and a single SNP. We first combine the n trait vectors $Y_i^T = (Y_{i1}, \ldots, Y_{ik})$ by row such that the resulting response matrix Y is of dimension $n \times k$, and the jth column of Y corresponds to phenotype j while the ith row of Y corresponds to the multiple traits from the ith subject; W is an $n \times q$ design matrix for covariates (including a column of 1s for the intercept); $x = (x_1, ..., x_n)^T$ is an $n \times 1$ vector of genotype scores (i.e., the counts of the minor allele) for the SNP.

The mvLMM can be written as

$$Y = W\lambda + x\psi^T + G + E, G \sim MN_{n\times k}(0, K, V_g), E \sim MN_{n\times k}(0, I_{n\times n}, V_\epsilon), \qquad (2.7)$$

where λ is a $q \times k$ matrix of regression coefficients for covariates; ψ is a $k \times 1$ vector of the SNP effect sizes for the k phenotypes; G is an $n \times k$ matrix of random effects; E is an $n \times k$ matrix of random errors; K is an $n \times n$ known kinship matrix, or more generally, a genetic relatedness matrix (GRM) estimated from whole-genome genotype data; $I_{n\times n}$ is an $n \times n$ identity matrix; V_g is a $k \times k$ symmetric matrix of genetic variance components; V_ϵ is a $k \times k$ symmetric matrix of environmental variance components; and $MN_{nk}(0, V_1, V_2)$ denotes the $n \times k$ matrix normal distribution with mean 0, a column covariance matrix V_1 of dimension $n \times n$, and a row covariance matrix V_2 of dimension $k \times k$. The goal is to test $H_0 : \psi = 0$.

A mvLMM differs from a standard LMM in that an $n \times n$ matrix K is used to account for possible genetic relatedness among all the subjects. Because the kinship matrix K may be full and may not be block diagonal, it means that all the subjects may be possibly correlated. However, as discussed by Zhou and Stephen [2014], the mvLMM can be rewritten more like a standard LMM in the following way. An eigendecomposition of the relatedness matrix K can be performed as $K = U_k D_k U_k^T$, where U_k is a $n \times n$ matrix of eigenvectors and D_k is a diagonal $n \times n$ matrix with diagonal elements corresponding to eigenvalues (i.e., $\text{diag}(\delta_1, \ldots, \delta_n)$). Then one can obtain the transformed phenotype matrix $Y = U_k Y$, transformed covariate matrix $W = U_k W$, transformed SNP vector $\tilde{x} = U_k x$, transformed random effect matrix $\tilde{G} = U_k G$, and transformed residual error matrix $\tilde{G} = U_k E$. After transformation, for each individual i, the transformed phenotypes given the transformed covariates and SNP follow independent (but not identical) multivariate normal distributions:

$$\tilde{y_i} = \lambda^T \tilde{w_i} + \psi \tilde{x_i} + \tilde{g_i} + \tilde{e_i}; \tilde{g_i} \sim MVN(0, \delta_i V_g), \tilde{e_i} \sim MVN(0, V_e), \tag{2.8}$$

where for $i = 1, \ldots, n$, $\tilde{y_i}^T$ is ith row vector of $\tilde{Y}$, $\tilde{w_i}^T$ is ith row vector of $\tilde{W}$, $\tilde{x_i}$ is ith element of vector $\tilde{x}$, $\tilde{g_i}^T$ is ith row vector of $\tilde{G}$ and $\tilde{e_i}^T$ is ith row vector of $\tilde{E}$; $Var(\tilde{Y_i}) = \delta_i V_g + V_e \equiv V_i$.

Based on model (2.8), one can write down the score vector:

$$U = \sum_{i=1}^{n} (\tilde{W_i}, \tilde{X_i})^T V_{i,0}^{-1} (\tilde{y_i} - \hat{\lambda}_0^T \tilde{w_i}), \tag{2.9}$$

where $\hat{\lambda}_0$ and $\hat{V}_{i,0}$ are obtained by fitting the null model under $H_0 : \tilde{y}_i = \lambda^T \tilde{w}_i + \tilde{g}_i + \tilde{e}_i$.

It is quite challenging to develop a fast algorithm to fit a mvLMM. Now such an algorithm is implemented in software package `GEMMA` [Zhou and Stephens, 2014]. However, as for most software packages, one is not able to obtain the score vector directly from the output. A simple and practical way to obtain the score vector is, as proposed earlier, to approximate it by the MLE and its covariance estimate, both available directly from the output of `GEMMA`; accordingly a score-based test can be simply constructed and applied.

Generalized linear mixed model

In a familial study, we observe that in each family i, subject j has a univariate trait Y_{ij}, q covariates $W_{ij} = (W_{ij1}, \ldots, W_{ijq})^T$ and p SNPs $X_{ij} = (X_{ij1}, \ldots, X_{ijp})^T$.We would like to test for association between the trait and the SNPs through a GLMM:

$$g(\mu_{ij}) = W_{ij}^T \lambda + X_{ij}^T \psi + b_i, b_i \sim N(0, \sigma_b^2), \tag{2.10}$$

where $g()$ is a link function, $\mu_{ij} = E(Y_{ij}|X_{ij}, W_{ij}, b_i)$ is the conditional mean of the trait for subject j in family i, $\lambda = (\lambda_1, \ldots, \lambda_q)^T$ is a $q \times 1$ vector of regression coefficients for covariates W_{ij}, $\psi = (\psi_1, \ldots, \psi_p)^T$ a $p \times 1$ vector of regression coefficients for SNP set X_{ij} , and b_i is a random effect inducing correlations among the traits of the subjects from the same family. The goal is to test $H_0 : \psi = 0$. However, in general, due to the lack of the closed form for the marginal likelihood, there is no closed-form expression for the score vector for ψ either [Breslow and Clayton, 1993]. Hence, it is not easy to develop a score-based test for such a model. Below we propose a new method to approximate the score vector, based on which it is straightforward to construct a score-based test.

2.2.3 New Method: Approximating the Score Vector

Estimation via Maximum Likelihood

Suppose we would like to test $H_0 : \psi = \psi_0$ in the presence of nuisance parameter λ. Denote $\hat{\theta}_0 = (\psi_0^T, \hat{\lambda}_0^T)^T$ as the restricted MLE of $\theta = (\psi^T, \lambda^T)^T$ under H_0, while $\hat{\theta} = (\hat{\psi}^T, \hat{\lambda}^T)^T$ as the unrestricted MLE of $\theta = (\psi^T, \lambda^T)^T$ (i.e., under H_1). Partition the

Fisher's information matrix H accordingly as

$$H = \begin{pmatrix} H_{\psi\psi} & H_{\psi\lambda} \\ H_{\lambda\psi} & H_{\lambda\lambda} \end{pmatrix}, H^{-1} = \begin{pmatrix} H^{\psi\psi} & H^{\psi\lambda} \\ H^{\lambda\psi} & H^{\lambda\lambda} \end{pmatrix}, \tag{2.11}$$

Denote the whole score vector for θ as $U_\theta(\theta) = (U_\psi(\theta)^T, U_\lambda(\theta)^T)^T$. As shown by Kent [1982, the equation following (4.1)],

$$U \equiv U_\psi(\widehat{\theta}_0) = (H^{\psi\psi})^{-1}(\hat{\psi} - \psi_0) + o_p(1). \tag{2.12}$$

Because the consistent estimator $\widehat{\mathrm{Cov}}(\hat{\psi}) = H^{\hat{\psi}\hat{\psi}}$, we have

$$U \approx \hat{\mathrm{Cov}}(\hat{\psi})^{-1}(\hat{\psi} - \psi_0), \widehat{\mathrm{Cov}}(U) = \widehat{\mathrm{Cov}}(\hat{\psi})^{-1}. \tag{2.13}$$

Thus, we first fit a full model (under H_1) to obtain the MLE $\hat{\psi}$ and its covariance estimate $\widehat{\mathrm{Cov}}(\hat{\psi})$, then we can approximate the score vector U and its covariance matrix accordingly. In this way, we can construct (approximate) score-based tests such as the score test, the SPU, and aSPU tests. In particular, it is easy to see that the approximate score-based score test is the same as the Wald test:

$$U^T \widehat{\mathrm{Cov}}(U)^{-1} U = (\hat{\psi} - \psi_0)^T \widehat{\mathrm{Cov}}(\hat{\psi})^{-1}(\hat{\psi} - \psi_0). \tag{2.14}$$

Estimation via Estimating Functions

For estimating function-based approaches, although a generalized score test [e.g., Boos, 1992; Kent, 1982; Rotnitzky and Jewell, 1990] can be constructed, the popular statistical software may not provide direct output of such (generalized) score vectors. For easy implementation, it may be useful to approximate the (generalized) score vector by the parameter estimate and its covariance matrix. Specifically, by treating an unbiased mean 0 estimating function as a (generalized) score function and by a Taylor expansion, we still have Equation (2.12). However, $\widehat{\mathrm{Cov}}(\hat{\psi})_M = H^{\hat{\psi}\hat{\psi}}$ is the model-based covariance estimator, which is not consistent unless all working assumptions hold (essentially assuming that the estimating function is indeed a score function). More generally, a consistent sandwich estimator $\widehat{\mathrm{Cov}}(\hat{\psi})_S$ is used. Hence, we can modify the score vector

approximation as

$$U \approx \widehat{\mathrm{Cov}}(\hat{\psi})_M^{-1}(\hat{\psi} - \psi_0), \mathrm{Cov}(U) \approx \widehat{\mathrm{Cov}}(\hat{\psi})_M^{-1}\widehat{\mathrm{Cov}}(\hat{\psi})_S\widehat{\mathrm{Cov}}(\hat{\psi})_M^{-1} \tag{2.15}$$

Accordingly, once we obtain the point estimate $\hat{\psi}$, its model-based covariance estimate $\widehat{\mathrm{Cov}}(\hat{\psi})_M$ and its sandwich estimate $\widehat{\mathrm{Cov}}(\hat{\psi})_S$, we can obtain an approximation to the score vector U, based on which we can construct a score-based test. Again it is easy to verify that the score test based on the approximate score is exactly the same as the Wald test.

We explored the use of such tests for marginal approaches to GLMMs for correlated binary data (i.e., GEE) [Liang and Zeger, 1986] (and to Cox regression for correlated survival data; not shown). Note that in general our proposed approximate score vector is derived based on an asymptotically normal point estimator, and thus is only asymptotically unbiased, while the generalized score vector is simply the estimating function being used and is often unbiased for finite samples; this difference leads to varying performances of an approximate score-based test and an exact generalized score test [Boos, 1992] for finite samples, though their difference diminishes as the sample size increases, as to be shown later for GEE.

2.3 Results

2.3.1 Example

About ADNI data

We applied the methods to the ADNI-1 data consisting of 681 non-Hispanic Caucasians with both genotypic and phenotypic data. The phenotypes were cortical thickness measures of some regions of interest (ROIs) in the brain; they were cross-sectionally processed using FreeSurfer by UCSF researchers [Hartig et al., 2012]. We tested on about 20 SNPs and several multivariate traits as considered in Shen et al. [2010] and Zhang et al. [2014]. For the purpose of illustration, we only show the results for two SNPs and four multivariate traits: APOE-4 in gene APOE that is well known to be

associated with AD, and rs7526034 on chromosome 1 (LOC199897), both were associated with multiple neuroimaging phenotypes [Shen et al., 2010]; the four multivariate traits were left and right sides of Par (denoted as LPar and RPar), each with four ROIs (inferior and superior parietal gyri, supramarginal gyrus, and precuneus), right side of Front with six ROIs (caudal midfrontal, rostral midfrontal, superior frontal, lateral orbitofrontal, medial orbitofrontal gyri, and frontal pole), right side of LatTemp with three ROIs (inferior temporal, middle temporal, and su- perior temporal gyri). Given a large number of parameters to be estimated ($> k^2$ with k traits) in a mvLMM, as pointed out by Zhou and Stephen [2014], only a small to moderate number of phenotypes (~ 210) were recommended to be used for a typical sample size for GWAS (i.e., n in thousands). Hence, with only a moderate sample size $n = 681$ here, we only considered a few multivariate traits containing no more than six univariate traits (otherwise, in addition to the questionable asymptotics, we also encountered some numerical convergence problems).

Analysis Results

As shown in Table 2.1, for most SNP-multivariate trait pairs, the aSPU test gave similar results as those of the classic Wald and Score tests. However, there were a few differences. Notably, for rs7526034-RLatTemp, the aSPU test gave a more significant P-value than those of the Wald and Score tests; on the other hand, for APOE-ϵ4-RPar, it was the reverse. Among the SPU tests, SPU(1) usually gave most significant results, presumably because of the smaller number of univariate traits (k) and the same direction of the SNP-univariate trait associations. In summary, our results demonstrate the feasibility of using our proposed method to approximate the score vector for a complex mvLMM, and accordingly construct the score-based aSPU test, which might be more powerful in some situations (to be shown in simulations) and thus can be complementary to the standard Wald and Score tests.

2.3.2 Simulations

Simulation I: mvLMM

To mimic real data, we used the ADNI data to generate multivariate phenotypes according to the fitted the mvLMM models (2.7) while using the covariates and SNPs in

the ADNI data too. Specifically, two SNP-phenotype pairs, rs7526034-RLatTemp and APOE-ϵ4-RPar were chosen; from their corresponding fitted models (2.7), we obtained the parameter estimates such as $\hat{\lambda}$, $\hat{\psi}$, $\hat{V}_g$, and $\hat{V}_\epsilon$.Those parameter estimates except $\hat{\psi}$, which was either 0 for the null model or was scaled by a factor 1/2 to reduce the effect size of the SNP for the non-null model (because we were using a nominal significance level at 0.05), were then used to simulate the phenotypes by model (3). For each simulated dataset, as before, the MLE of ψ and its covariance estimate were obtained from `GEMMA` to approximate its score vector so that the SPU and aSPU tests could be applied. We used $B = 1,000$ to calculate their P-values. As a comparison, we also used the Wald test, Score test, and LRT directly output by GEMMA. Based on 5,000 replicates for each setup, we obtained the empirical Type I error and power estimates. However, we note that there were some convergence problems when running GEMMA for about 1% and 2% of simulated datasets for the two SNP-phenotype pairs, respectively; our results were based on the remaining ones without any convergence problems.

As shown in Table 2.2, the Type I error rates for both pairs were only slightly inflated for all the tests except the LRT, which had largely inflated Type I error rates. The inflation could be due to a large number of parameters to be estimated in an mvLMM with a moderate sample size.

Because the Type I error rates based on fitting the mvLMM were slightly inflated, while it was known that there was barely any population stratification in the ADNI data [Xu et al., 2014], we fitted the corresponding model after treating $K = I$; however,we experienced some numerical convergence problems in fitting the mvLMM, likely due to that the two unstructured covariance matrices V_g and V_ϵ were not identifiable (after forcing $K = I$). Thus we simply used function `gls()` in R package `nlme` to fit a corresponding marginal linear model with or without top 10 principal components (PCs); the PCs were extracted using Plink [Purcell et al., 2007] based on almost a half million SNPs of the 757 subjects in the ADNI data. As shown in Table 2.3, the Type I error rates were better controlled. Note that because the simulated data were generated with $K \neq I$, some slight inflation of a Type I error rate was expected under the incorrect assumption $K = I$.

The two SNP-phenotype pairs were chosen partly because in the ADNI data analysis the aSPU test gave a more significant P-value than those of the Wald and Score tests

for pair 1, while it was the opposite for pair 2 (Table 2.1). It was confirmed that in the simulations the aSPU test was indeed slightly more (or less) powerful than the Wald and Score tests for pair 1 (or pair 2).

Simulation II: LMM

We considered a case with unrelated individuals and multiple quantitative traits similar to those in Zhang et al. [2014], for which the exact score vector could be derived. We would compare the performance of an exact score-based test with that of its approximate score-based one. For each subject $i = 1, \ldots, n$, we generated his/her genotype data as in Pan [2009]. Specifically, for each subject i, we first generated a latent vector $G_i = (G_{j1}, \ldots, G_{j,p+1})'$ from a multivariate normal distribution with a first-order autoregressive (AR-1) covariance structure with parameter $\rho = 0.5$: $\text{Cov}(G_{is}, G_{it}) = \rho^{|s-t|}$. Second, each latent element G_{is} was dichotomized to 0 or 1 with probability $\text{Prob}(G_{is} = 1)$ as its minor allele frequency (MAF), randomly drawn from a uniform distribution. Third, we independently generated another haplotype for subject i, then combined the two haplotypes to form the genotypes for subject j . In this way, we obtained the genotypes of all the subjects.

The first SNP was chosen as the causal one with MAF randomly drawn from a uniform distribution $U(0.3, 0.4)$, while the MAFs of the other SNPs were independently drawn from $U(0.1, 0.5)$. For each subject i, we simulated k traits $Y_i = (Y_{i1}, \ldots, Y_{ik})$ from a linear model:

$$Y_i = \lambda + x_i\psi + \epsilon_i, \tag{2.16}$$

where $\lambda = (\lambda_1, \ldots, \lambda_k)^T$, $\psi = (\psi_1, \ldots, \psi_k)^T$; $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2 R)$, with $\sigma = 1$ and R as a compound symmetry (CS) correlation matrix with the correlation $r = 0.3$; x_i is the genotype dosage of the causal SNP. Under H_0 we had $\psi = 0$; under H_1 we had $\psi_m \neq 0$ for $1 \leq m \leq 5$ and $\psi_m = 0$ for $5 < m \leq k$. The non-zero ψ_j's were simulated from a uniform distribution $U(0.2, 0.3)$. In this way, under H_1, only the first five traits were associated with the causal SNP, that is, as $k = 5, 10, 20, 30, 40$, we gradually increased the number of the nonassociated traits from 0 to 5, 15, 25 and 35.

The simulated data were fitted by a LMM:

$$Y_{ij} = \lambda_j + x_i\psi_j + b_i + \epsilon_{ij}, b_i \sim N(0, \sigma_b^2), \epsilon \sim N(0, \sigma_e^2), \tag{2.17}$$

where b_i was a normal random effect to model the correlations among multiple traits, ϵ_{ij} was the random error, x_i was a scalar corresponding to the genetic score of the SNP nearest to the causal SNP.

We implemented both the approximate score-based and exact score-based tests. For an approximate score-based test, we fitted model (2.17) and used the MLE of ψ to approximate its score vector and its variance-covariance matrix. For an exact score-based test, we fitted the reduced LMM modeled under H_0, $Y_{ij} = \lambda_j + b_i + \epsilon_{ij}$, to obtain the MLEs $\hat{\lambda} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_k)^T$. Denote I as the $k \times k$ identity matrix. The exact score vector and its variance-covariance matrix can be written as

$$U = \sum_{i=1}^{n} (I, X_i)^T \hat{V}_i^{-1}(Y_i - \hat{\lambda}), \mathrm{Cov}(U) = \sum_{i=1}^{n} (I, X_i)^T \hat{V}_i^{-1}(I, X_i), \tag{2.18}$$

where $\hat{V}_i$ was the MLE of V_i with its diagonal elements $\hat{\sigma}_b^2 + \hat{\sigma}_\epsilon^2$ and off-diagonal elements $\hat{\sigma}_b^2$. Partition the score vector and its covariance into two parts corresponding to the interceptand ψ parameters, respectively,

$$U = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix}, \mathrm{Cov}(U) = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}, \tag{2.19}$$

then we have the exact score vector for ψ as U_2 and its covariance matrix $\mathrm{Cov}(U_2) = V_{22} - V_{21}V_{11}^{-1}V_{12}$.

To estimate the Type I error and power, $1,000$ datasets were independently simulated and analyzed. Each of the $1,000$ datasets consisted of $1,000$ subjects. We used $B = 1,000$ to obtain P-values for any permutation based methods. As a comparison, we also showed the results from the UminP.

As shown in Table 2.4, first, regardless of the test being examined, its version based on the approximate score vector and that based on the exact score vector gave almost the same results, suggesting the high accuracy of the asymptotic approximation in this case. Second, we note that the Type I error rates were satisfactorily controlled, even

for 40 traits. Third, in agreement with Zhang et al. [2014], the aSPU test was more powerful than the score test for five traits, but not for other numbers of traits; both were much more powerful than single trait based UminP test, presumably due to the former twos combining information across the five associated traits.

Simulation III: GLMM

We considered a familial/trio study design with a single binary trait; because there were multiple subjects in each family, their traits might be correlated. For each of the two parents in each family $i = 1, \ldots, n$, we generated their haplotypes and thus genotypes as described in the previous section (with $\rho = 0.8$); then their offsprings haplotype and thus genotype data were obtained according to the Mendelian transmission. In this way, we obtained the genotype data with $p+1$ SNPs for each subjects. The SNP at the center (i.e., at position $p/2 + 1$) was chosen as the causal one with MAF 0.3, while the MAFs of the other SNPs were independently drawn from $U(0.1, 0.4)$. Figure 1 shows linkage disequilibrium plots for the generated SNPs (after the causal SNP was removed) based on one parent from each family.

For subject j in family i, denote x_{ij} as the genotypic score for the causal SNP, and $X_{ij} = (X_{ij1}, \ldots, X_{ijp})^T$ as a vector of the genotypic scores for the p noncausal SNPs. The disease indicator $Y_{ij} = 0$ or 1 was generated from the below GLMM: Assuming λ_0 is the background log odds ratio, ψ_0 is the effect sizes, the resulting GLMM model can be written as

$$\text{Logit}(E[Y_{ij}|b_i]) = \lambda_0 + x_{ij}\psi_0 + b_i, b_i \sim N(0, \sigma_b^2), \tag{2.20}$$

where $\lambda_0 = -\log(4)$ was chosen to have a 20% background disease prevalence, while ψ_0 at varying effect sizes was used to investigate the empirical Type I error ($\psi_0 = \log(1)$) and power ($\psi_0 > \log(1)$). We fixed $\sigma_b = 1$, and considered two cases with ($p = 10$, $n = 200$) and ($p = 20$, $n = 400$). For each simulation setup, we generated 1,000 simulated datasets to estimate the empirical Type I error or power.

We fitted both a GLMM (2.10) and a corresponding marginal model to test $H_0 : \psi = 0$. A GLMM was fitted using function `glmer()` from R package `lme4`. Either the Laplace approximation (LA) or adaptive Gaussian quadrature (AGQ) was used to approximate the (marginal) log-likelihood. For AGQ, we specified the number of points

per axis nAGQ = 25; in the manual of `lme4`, it was mentioned that "a model with a single, scalar random-effects term could reasonably use up to 25 quadrature points per scalar integral" [Bates et al., 2014]. For a marginal model, both a working independence and a working CS correlation structures were used in GEE. We then applied our method to approximate the score vector for the fitted GLMM and GEE models, respectively, based on which we applied the SPU and aSPU tests; we used B = 1, 000 to calculate their P-values. As a comparison, we also showed the results from the UminP and Wald tests; recall that the Wald test is equivalent to the approximate score test.

As shown in Table 2.5, it seems that the Type I error rates were appropriately controlled by all the tests except the Wald test, which gave slightly inflated Type I error rates in GEE as well known in the literature [e.g., Zhang et al., 2014]. It is clear that, for the same fitted model, the aSPU test was much more powerful than the popular UminP and Wald tests. Among the SPU tests, the SPU(1) was nearly as powerful as SPU(2) and SPU(3) for a smaller number of parameters to be tested with $p = 10$, but it was less powerful than the latter two for $p = 20$. This is in agreement with the analysis and motivation of the SPU tests: for p = 10, because all the 10 SNPs were correlated with the causal SNP, the SPU(1) test was expected to be powerful, as more generally known for the burden tests; on the other hand, for $p = 20$, because some SNPs were barely correlated with the causal SNP, to minimize the effects on power of the nonassociated SNPs, a larger γ (here $\gamma = 3$) would yield higher power for the SPU(γ) test. Among the three fitted models, the aSPU test based on GEE(CS) was more powerful than that based on the other two models, the latter of which gave similar results; the lower power of GEE(Ind) might be due to the use of the working independence correlation structure, while the GLMM was fitted by an approximate maximum likelihood (using either the LA by default, or ACQ in the function glme), leading to loss of efficiency. It is noted that the results from the LA and ACQ approximations were very close; they seemed to be a little conservative with Type I rates much lower than the nominal 0.05.

Because the proposed method of approximating the score vector is asymptotic, to further evaluate its finite-sample performance, we also applied the score-based tests based on the exact, not approximate, GEE (generalized) score formulas, as implemented in Zhang et al. [2014]. Comparing the test results in Tables 2.5 and 2.6, we can see that the approximate score-based tests had slight losses of power for the smaller sample size

$n = 200$, but performed equally well as the exact score-based tests for the larger sample size $n = 400$.

2.4 Conclusions

We have described an asymptotic approach to approximating the score vector in some complex models, such as a mvLMM or GLMM. The approximate score vector can then be used to construct any score-based tests, including KMR and aSPU tests for multiple traits or familial data. Using both real and simulated data, we have demonstrated such approximate score-based tests can improve power over (and control the Type I error rate better than) the standard Wald test and the UminP test that has been widely used in GWAS. The proposed approximate score vector offers a simple and general way to extend many score-based tests to other complex models, in which the score vector is unavailable from the statistical package being used. Although we have focused on the mvLMM and GLMM, we also considered the LMM, the Cox frailty model [Therneau et al., 2003], and the Cox mixed-effects model [Therneau, 2012], and reached similar conclusions (results not shown to save space); the difference between the two Cox models is that the latter includes a random effect to account for genetic relatedness across all subjects, similar to that adopted in the mvLMM. We envision the use of the proposed approximate score vector in other models.

Table 2.1: P-values of the various mvLMM-based tests in analysis of the ADNI data.

SNP	Test	LPar(4)	RFront(6)	RLatTemp(3)	RPar(4)
APOE-ϵ4	SPU(1)	1.2×10^{-5}	2.0×10^{-7}	2.0×10^{-5}	1.3×10^{-6}
	SPU(2)	8.3×10^{-2}	2.4×10^{-2}	1.3×10^{-1}	3.8×10^{-4}
	SPU(3)	1.9×10^{-1}	2.4×10^{-3}	1.9×10^{-2}	1.3×10^{-2}
	SPU(4)	1.6×10^{-1}	7.7×10^{-3}	6.8×10^{-2}	1.1×10^{-3}
	SPU(5)	3.6×10^{-1}	4.3×10^{-3}	3.5×10^{-2}	2.1×10^{-2}
	SPU(∞)	2.9×10^{-1}	5.2×10^{-3}	4.8×10^{-2}	7.8×10^{-3}
	*a*SPU	3.5×10^{-5}	5.0×10^{-7}	6.0×10^{-5}	3.8×10^{-6}
	Wald	1.2×10^{-5}	1.4×10^{-5}	4.6×10^{-5}	6.0×10^{-8}
	score	1.8×10^{-5}	2.3×10^{-5}	5.9×10^{-5}	1.4×10^{-7}
rs7526034	SPU(1)	5.7×10^{-2}	9.0×10^{-4}	$<1.0\times10^{-7}$	6.2×10^{-2}
	SPU(2)	8.2×10^{-1}	5.7×10^{-2}	2.0×10^{-1}	1.4×10^{-2}
	SPU(3)	6.5×10^{-1}	5.8×10^{-2}	6.8×10^{-2}	4.8×10^{-2}
	SPU(4)	8.4×10^{-1}	8.3×10^{-2}	2.1×10^{-1}	2.6×10^{-2}
	SPU(5)	7.7×10^{-1}	7.5×10^{-2}	1.5×10^{-1}	4.8×10^{-2}
	SPU(∞)	8.4×10^{-1}	8.4×10^{-2}	2.7×10^{-1}	4.0×10^{-2}
	*a*SPU	1.3×10^{-1}	2.3×10^{-3}	$<1.0\times10^{-7}$	3.5×10^{-2}
	Wald	2.6×10^{-1}	1.4×10^{-5}	2.0×10^{-6}	3.5×10^{-4}
	score	2.6×10^{-1}	1.6×10^{-5}	3.1×10^{-6}	4.3×10^{-4}

Table 2.2: Simulation I: empirical Type I error and power for two SNP-phenotype pairs, rs7526034-RLatTemp (pair 1) and APOE-ϵ4-RPar (pair 2), based on fitting mvLMM.

		Approximate score vector									
		SPU(γ)									
Pair		$\gamma=1$	2	3	4	5	∞	aSPU	Wald	Score	LRT
1	Type I	0.062	0.067	0.067	0.066	0.069	0.068	0.066	0.067	0.067	0.178
	Power	0.722	0.103	0.195	0.112	0.139	0.120	0.621	0.599	0.591	0.604
2	Type I	0.058	0.065	0.063	0.066	0.065	0.067	0.068	0.064	0.058	0.188
	Power	0.644	0.389	0.335	0.381	0.336	0.361	0.643	0.689	0.674	0.641

Table 2.3: Simulation I: empirical Type I error and power for two SNP-phenotype pairs, rs7526034-RLatTemp (pair 1) and APOE-ϵ4-RPar (pair 2) based on fitting marginal linear models (LMs) with or without the top 10 PCs.

			Approximate score							
			SPU(γ)							
Model	Pair		$\gamma = 1$	2	3	4	5	∞	aSPU	Wald
LM, 10 PCs	1	Type I	0.049	0.057	0.063	0.057	0.061	0.058	0.061	0.059
		Power	0.742	0.092	0.212	0.117	0.158	0.134	0.648	0.611
	2	Type I	0.050	0.052	0.050	0.051	0.048	0.049	0.052	0.052
		Power	0.616	0.347	0.306	0.347	0.316	0.339	0.610	0.664
LM, no PCs	1	Type I	0.049	0.059	0.062	0.057	0.061	0.057	0.060	0.057
		Power	0.749	0.095	0.214	0.116	0.160	0.136	0.645	0.609
	2	Type I	0.052	0.056	0.054	0.057	0.052	0.056	0.054	0.053
		Power	0.633	0.360	0.317	0.361	0.327	0.352	0.634	0.683

Table 2.4: Simulation II: empirical type I error rates and power of the approximate (approx) score- and exact score-based tests when multiple traits were correlated with a CS structure with correlation $r = 0.3$.

				SPU(γ)							
	Score vector	No. of traits	UminP	$\gamma = 1$	2	3	4	5	∞	aSPU	Score
Type I	approx	5	0.041	0.042	0.040	0.048	0.038	0.044	0.043	0.042	0.034
		10	0.050	0.051	0.056	0.056	0.057	0.059	0.053	0.052	0.059
		20	0.048	0.057	0.046	0.048	0.049	0.047	0.050	0.045	0.047
		30	0.048	0.048	0.051	0.050	0.049	0.051	0.050	0.053	0.053
		40	0.047	0.052	0.042	0.046	0.048	0.045	0.051	0.042	0.046
	exact	5	0.043	0.045	0.034	0.042	0.038	0.043	0.043	0.039	0.034
		10	0.052	0.051	0.056	0.057	0.054	0.052	0.050	0.052	0.059
		20	0.046	0.053	0.051	0.050	0.049	0.048	0.049	0.046	0.047
		30	0.047	0.047	0.053	0.056	0.045	0.053	0.050	0.053	0.053
		40	0.051	0.055	0.039	0.047	0.045	0.042	0.049	0.046	0.046
Power	approx	5	0.140	0.686	0.139	0.288	0.146	0.188	0.144	0.549	0.435
		10	0.324	0.274	0.528	0.341	0.454	0.331	0.324	0.486	0.590
		20	0.394	0.097	0.597	0.433	0.544	0.403	0.392	0.512	0.606
		30	0.378	0.088	0.581	0.395	0.540	0.399	0.375	0.493	0.570
		40	0.364	0.071	0.538	0.416	0.532	0.404	0.368	0.474	0.543
	exact	5	0.138	0.685	0.133	0.285	0.141	0.182	0.141	0.543	0.434
		10	0.323	0.274	0.536	0.334	0.458	0.315	0.322	0.479	0.588
		20	0.384	0.099	0.594	0.426	0.544	0.401	0.397	0.519	0.606
		30	0.384	0.080	0.571	0.399	0.543	0.398	0.378	0.500	0.569
		40	0.369	0.068	0.534	0.419	0.531	0.410	0.368	0.470	0.542

The first five traits were associated with a causal SNP with effect size $\beta_j \sim U(0.2, 0.3)$; the SNP nearest to the causal SNP was tested.

Table 2.5: Simulation III: empirical Type I error (for OR = 1) and power (for OR > 1) of the approximate score-based tests with correlated binary traits.

				SPU(γ)							
Model	Case	OR	UminP	$\gamma=1$	2	3	4	5	∞	aSPU	Wald
GLMM	1	1	0.018	0.044	0.015	0.016	0.011	0.012	0.014	0.020	0.021
(LA)		1.4	0.078	0.125	0.100	0.105	0.088	0.090	0.074	0.100	0.057
		1.8	0.216	0.323	0.296	0.294	0.266	0.257	0.195	0.300	0.141
		2.2	0.406	0.514	0.494	0.511	0.461	0.452	0.368	0.470	0.280
		2.6	0.580	0.702	0.681	0.680	0.645	0.641	0.568	0.667	0.427
		3	0.736	0.825	0.822	0.826	0.799	0.794	0.713	0.807	0.587
	2	1	0.028	0.030	0.018	0.024	0.017	0.019	0.024	0.023	0.032
		1.4	0.096	0.157	0.116	0.138	0.106	0.108	0.090	0.131	0.067
		1.8	0.344	0.424	0.422	0.467	0.425	0.431	0.353	0.414	0.187
		2.2	0.677	0.668	0.750	0.776	0.750	0.755	0.667	0.748	0.419
		2.6	0.862	0.834	0.897	0.914	0.902	0.904	0.848	0.893	0.666
		3	0.950	0.918	0.964	0.969	0.968	0.972	0.938	0.963	0.825
GLMM	1	1	0.015	0.047	0.014	0.011	0.008	0.010	0.011	0.022	0.018
(AGQ)		1.4	0.073	0.118	0.084	0.096	0.085	0.083	0.070	0.092	0.039
		1.8	0.198	0.310	0.279	0.281	0.248	0.242	0.187	0.281	0.122
		2.2	0.390	0.505	0.472	0.489	0.439	0.431	0.353	0.463	0.255
		2.6	0.565	0.688	0.669	0.664	0.624	0.616	0.536	0.654	0.397
		3	0.724	0.816	0.817	0.817	0.792	0.783	0.695	0.794	0.557
	2	1	0.024	0.029	0.016	0.023	0.016	0.015	0.022	0.024	0.025
		1.4	0.085	0.149	0.110	0.133	0.096	0.103	0.083	0.122	0.056
		1.8	0.329	0.407	0.408	0.450	0.396	0.415	0.332	0.412	0.166
		2.2	0.665	0.667	0.727	0.764	0.730	0.737	0.657	0.744	0.374
		2.6	0.851	0.833	0.892	0.910	0.895	0.901	0.834	0.888	0.637
		3	0.947	0.916	0.962	0.970	0.968	0.970	0.928	0.958	0.810
GEE	1	1	0.030	0.040	0.024	0.023	0.026	0.024	0.026	0.032	0.057
(Ind)		1.4	0.093	0.134	0.108	0.111	0.100	0.099	0.088	0.116	0.106
		1.8	0.215	0.291	0.284	0.288	0.253	0.248	0.202	0.272	0.170
		2.2	0.375	0.478	0.459	0.467	0.428	0.422	0.366	0.445	0.308
		2.6	0.519	0.625	0.625	0.628	0.596	0.588	0.519	0.603	0.438
		3	0.672	0.754	0.755	0.757	0.736	0.729	0.656	0.740	0.558
	2	1	0.036	0.037	0.035	0.033	0.032	0.034	0.042	0.033	0.081
		1.4	0.113	0.146	0.128	0.149	0.134	0.137	0.110	0.146	0.125
		1.8	0.323	0.368	0.368	0.410	0.377	0.394	0.320	0.385	0.229
		2.2	0.589	0.593	0.662	0.709	0.677	0.680	0.579	0.672	0.427
		2.6	0.775	0.761	0.829	0.847	0.826	0.830	0.770	0.825	0.619
		3	0.902	0.865	0.926	0.939	0.932	0.928	0.881	0.925	0.755
GEE	1	1	0.044	0.061	0.041	0.037	0.035	0.035	0.034	0.040	0.065
(CS)		1.4	0.116	0.141	0.133	0.136	0.115	0.117	0.111	0.133	0.104
		1.8	0.271	0.366	0.349	0.358	0.323	0.319	0.261	0.337	0.222
		2.2	0.459	0.576	0.573	0.577	0.541	0.538	0.459	0.547	0.385
		2.6	0.653	0.730	0.734	0.740	0.704	0.698	0.632	0.738	0.526
		3	0.800	0.845	0.862	0.862	0.847	0.840	0.780	0.851	0.670
	2	1	0.039	0.041	0.032	0.035	0.035	0.038	0.041	0.038	0.070
		1.4	0.124	0.183	0.159	0.180	0.151	0.163	0.124	0.167	0.137
		1.8	0.433	0.476	0.500	0.543	0.509	0.508	0.438	0.506	0.309
		2.2	0.750	0.712	0.805	0.832	0.806	0.809	0.755	0.791	0.544
		2.6	0.895	0.881	0.925	0.943	0.933	0.931	0.883	0.925	0.749
		3	0.969	0.934	0.971	0.982	0.978	0.978	0.962	0.976	0.881

The two cases were for ($n = 200$, $p = 10$) and ($n = 400$, $p = 20$). A GLMM was fitted using either the Laplace (LA) or adaptive Gaussian quadrature (AGQ) approximation; the working correlation structure in GEE was assumed to be either independent(Ind) or compound symmetry (CS).

Table 2.6: Simulation III: empirical Type I error (for OR = 1) and power (for OR > 1) of the GEE score-based tests with correlated binary traits.

				SPU(γ)										
Model	Case	OR	UminP	$\gamma = 1$	2	3	4	5	6	7	8	∞	aSPU	Score
GEE	1	1	0.036	0.047	0.037	0.038	0.031	0.027	0.031	0.033	0.033	0.034	0.038	0.042
(Ind)		1.4	0.090	0.129	0.120	0.117	0.108	0.106	0.099	0.097	0.093	0.090	0.117	0.074
		1.8	0.228	0.295	0.305	0.303	0.277	0.264	0.254	0.248	0.245	0.220	0.284	0.148
		2.2	0.404	0.474	0.480	0.489	0.459	0.450	0.433	0.428	0.425	0.399	0.462	0.280
		2.6	0.567	0.630	0.654	0.658	0.623	0.614	0.603	0.591	0.577	0.555	0.625	0.406
		3	0.698	0.766	0.786	0.779	0.759	0.752	0.740	0.736	0.727	0.694	0.765	0.544
	2	1	0.049	0.046	0.044	0.038	0.039	0.042	0.045	0.044	0.044	0.048	0.051	0.054
		1.4	0.124	0.141	0.152	0.158	0.148	0.151	0.137	0.134	0.125	0.115	0.153	0.092
		1.8	0.355	0.369	0.403	0.429	0.413	0.415	0.392	0.385	0.368	0.344	0.413	0.198
		2.2	0.627	0.595	0.690	0.725	0.696	0.699	0.672	0.669	0.658	0.613	0.693	0.399
		2.6	0.803	0.770	0.852	0.865	0.843	0.841	0.827	0.825	0.820	0.787	0.841	0.597
		3	0.915	0.875	0.941	0.949	0.941	0.942	0.933	0.930	0.924	0.895	0.934	0.745
GEE	1	1	0.040	0.066	0.047	0.048	0.044	0.043	0.042	0.041	0.041	0.043	0.050	0.041
(CS)		1.4	0.119	0.140	0.135	0.134	0.127	0.122	0.120	0.116	0.114	0.114	0.130	0.082
		1.8	0.282	0.361	0.370	0.369	0.342	0.329	0.315	0.308	0.300	0.276	0.349	0.192
		2.2	0.487	0.572	0.582	0.585	0.554	0.546	0.533	0.525	0.510	0.474	0.554	0.357
		2.6	0.651	0.738	0.763	0.747	0.727	0.719	0.700	0.693	0.688	0.647	0.737	0.506
		3	0.823	0.847	0.865	0.870	0.850	0.845	0.837	0.834	0.829	0.797	0.865	0.644
	2	1	0.047	0.044	0.041	0.040	0.045	0.048	0.038	0.041	0.039	0.044	0.049	0.054
		1.4	0.150	0.172	0.162	0.183	0.159	0.168	0.154	0.159	0.149	0.140	0.168	0.098
		1.8	0.446	0.465	0.519	0.543	0.527	0.525	0.502	0.502	0.486	0.465	0.512	0.258
		2.2	0.766	0.712	0.825	0.832	0.823	0.823	0.805	0.807	0.793	0.770	0.816	0.507
		2.6	0.908	0.879	0.928	0.947	0.937	0.942	0.931	0.928	0.922	0.896	0.931	0.739
		3	0.973	0.936	0.975	0.982	0.981	0.979	0.979	0.979	0.978	0.967	0.978	0.866

The two cases were for ($n = 200$, $p = 10$) and ($n = 400$, $p = 20$).

Chapter 3

Binomial mixture model based association testing to account for genetic heterogeneity for GWAS

3.1 Introduction

The main part of this chapter has been published in Xu and Pan [2016].

Genome-wide association studies (GWAS) have been extremely successful in identifying thousands of common genetic variants, mostly single-nucleotide polymorphisms (SNPs), associated with common disease and complex traits (NHGRI Catalog: `http://wwww.genome.gov/gwastudies/`). Some important discoveries on the genetic architecture for complex traits are the following. First, genetic heterogeneity is everywhere: multiple SNPs and risk loci have been identi-fied for many common disease and complex traits, suggesting the plausibility and even ubiquity of a polygenic model. Second, the effect sizes of most causal SNPs are estimated to be from moderate to small, many of which with smaller effect sizes are yet to be identified. Consequently, larger sample sizes and more powerful statistical tests are always needed in order to identify more risk loci. However, almost all GWAS have adopted single SNP based analysis without explicitly accounting for (unknown) genetic heterogeneity, leading to possible loss of power as convincingly shown in simulation studies [Londono et al., 2012; Qian and Shao, 2013;

Zhou and Pan, 2009]. Note we consider unknown genetic heterogeneity here, rather than known phenotype heterogeneity as discussed in Darabi and Humphreys [2011]. Based on a two-component binomial mixture model of Zhou and Pan [2009], if any given locus contains a causal SNP, then the patient population is decomposed into two subpopulations: the first subpopulation consists of the patients whose disease is associated with the disease allele at the locus, whereas the second includes those with disease caused by other unknown alleles at other unknown risk loci. In addition to contrasting the allele frequencies (i.e., means or first moments) between the control and case groups as targeted by the most popular 1 df trend test, the proposed mixture model can capture some other distributional differences of the allele (i.e., second moments) between the two groups. For example, in an extreme case, even if there is barely any difference of the allele frequencies between the two groups, leading to no power of the trend test, if the mixture model assumption holds, then a mixture model based likelihood ratio test (LRT-H) may still be able to detect the distributional differences of the allele between the two groups. Qian and Shao [2013] extended the two-component binomial mixture model to one with more than two components, for which an LRT-H statistic with a simple closed-form and an asymptotic null distribution was derived, facilitating its application to GWAS. Surprisingly, to our knowledge, it has not yet been applied to any GWAS. In fact, we are not aware of any other analyses of GWAS that explicitly account for genetic heterogeneity. Here we review the LRT-H and apply it to the Wellcome Trust Case Control Consortium (WTCCC) GWAS data. To ensure that our conclusion is not limited to any specific disease, we considered multiple diseases. We demonstrate that the LRT-H test can be much more powerful than the popular trend test in identifying a much larger number of associated SNPs and risk loci.

3.2 Methods

For each subject i, suppose $X_i = 0$, 1 or 2 is the number of the minor allele of a SNP to be tested, and $Y_i = 0$ or 1 is the disease indicator. The methods are all based on single SNP analysis by testing on each SNP individually and separately, hence we can focus on only one SNP. The goal is to test for possible association between the SNP and disease.

3.2.1 The Trend Test and Related Tests

Most of the existing association tests ignore possible genetic heterogeneity due to the disease. For example, the most popular Cochran-Armitage 1df trend test can be formulated as the Score test in a logistic regression model [Wellek and Ziegler, 2011] (or more generally a GLM or Cox PHM for other types of traits)

$$\text{logit}(\Pr(Y_i = 1)) = \beta_0 + \beta_1 X_i \tag{3.1}$$

to test the null hypothesis H_0: $\beta_1 = 0$. It is well known [Clayton et al., 2004] that the Score test is

$$T_S = \frac{\bar{X}^{(1)} - \bar{X}^{(0)}}{\sqrt{\widehat{\text{Var}}(\bar{X}^{(1)} - \bar{X}^{(0)})}},$$

where $\bar{X}^{(d)}$ is the sample mean of X_i's with $Y_i = d$ for $d = 0$ or 1. The Score test is asymptotically equivalent to the Z-test for one SNP (and equivalent to Hotelling's T^2 test for multiple SNPs [Xiong et al., 2002; Fan and Knapp, 2003]). Specifically, one can model the conditional distribution of X_i as binomial:

$$(X_i|Y_i = 0) \sim Bin(2, \theta_0), \qquad (X_i|Y_i = 1) \sim Bin(2, \theta^*),$$

for which we'd like to test the null hypothesis H_0': $\theta_0 = \theta^*$, which is equivalent to the original H_0. It is easy to see that, the Wald test for H_0' is

$$T_W = \frac{\bar{X}^{(1)} - \bar{X}^{(0)}}{\sqrt{\widetilde{\text{Var}}(\bar{X}^{(1)} - \bar{X}^{(0)})}},$$

which differs from T_S in the variance estimates used in the denominator, but nonetheless is asymptotically equivalent to T_S. Furthermore, the asymptotically equivalent likelihood ratio test (LRT) can be also applied:

$$T_L = 2\log(L_H L_D) - 2\log L_0$$

with

$$L_H = \prod_{g=0}^{2} B_2(g, \hat{\theta}_0)^{m_g}, \quad L_D = \prod_{g=0}^{2} B_2(g, \hat{\theta}_1)^{n_g}, \quad L_0 = \prod_{g=0}^{2} B_2(g, \hat{\theta}_{01})^{m_g+n_g},$$

where n_g and m_g are the genotype frequencies as summarized in Table 3.1, the maximum likelihood estimates (MLEs) of the minor allele frequencies (MAFs) under H_0 and H_1 are

$$\hat{\theta}_{01} = \frac{2n_2 + 2m_2 + n_1 + m_1}{2n + 2m}, \quad \hat{\theta}_0 = \frac{2m_2 + m_1}{2m}, \quad \hat{\theta}_1 = \frac{2n_2 + n_1}{2n},$$

and

$$B_2(g, p) = \Pr(X = g) = \binom{2}{g} p^g (1-p)^{2-g}$$

is the probability mass function for a binomial distribution $X \sim Bin(2, p)$.

The above three tests all share the same (asymptotic) null distribution as a chi-squared distribution χ^2_1 with 1 df.

Rather than using a trend test based on the additive genetic model for X_i, a more general 2df test can be formulated by fitting an expanded regression model:

$$\text{logit}(\Pr(Y_i = 1)) = \beta_0 + \beta_1 X_i + \beta_2 X_i^2, \tag{3.2}$$

and we test $H_0'' : \beta_1 = \beta_2 = 0$ with one of the three asymptotically equivalent Score test, Wald test and LRT, all with an asymptotically null distribution of χ^2_2 with df=2. Interestingly, as pointed out by Kim et al. [2010], β_2 measures the difference of Hardy-Weinberg coefficients in the disease and control groups, hence a 2df test can be regarded as testing on both allele frequency difference and Hardy-Weinberg coefficient difference between the two groups. In this chapter, we used R function `glm()` to fit a logistic regression model and applied the Wald test.

3.2.2 Hardy-Weinberg Equilibrium (HWE) Exact test

A Hardy-Weinberg Equilibrium (HWE) test can be applied to the case group (with notation shown in Table 3.1) for association analysis [Nielsen et al., 1998]. Given that the total number of observed minor allele is $n_a = 2n_2 + n_1$, under the assumption of

HWE, the probability of observing n_1 heterozygotes:

$$P(N_1 = n_1|n, n_a) = \frac{2^{n_1} n!}{n_2! n_1! n_0!} \times \frac{n_a!(2n - n_a)!}{(2n)!}, \tag{3.3}$$

and the p-value is calculated as

$$P_{HWE} = \sum_{n_1^*} I[P(N_1 = n_1|n, n_a) \geq P(N_1 = n_1^*|n, n_a)] \times P(N_1 = n_1^*|n, n_a). \tag{3.4}$$

We conducted the HWE exact test of Wigginton et al. [2005] as implemented in function `hwexact()` from R package `hwde`.

3.2.3 Association Testing under Genetic Heterogeneity

To fully and explicitly account for genetic heterogeneity, Zhou and Pan (2009) proposed a binomial mixture model for the disease group while using a usual binomial model for the control group:

$$(X_i|Y_i = 0) \sim Bin(2, \theta_0), \qquad (X_i|Y_i = 1) \sim \pi Bin(2, \theta) + (1 - \pi)Bin(2, \theta_0), \tag{3.5}$$

where θ_0 is the background MAF for the controls. In contrast, for the case group, we assume θ is the probability of having the minor allele on a chromosome for a subpopulation of cases with disease caused by (or associated with) the minor allele, while for other subpopulations of cases the disease is caused by (unknown) variants at other unlinked loci, and thus for them the probability of having the minor allele at the locus of interest is the same as that for the controls. We test $H_0 : \theta = \theta_0$ or $\pi = 0$. Zhou and Pan [2009] considered more general scenarios with different θ_0's for cases and controls, or with more than one non-null component for cases, but recommended the above two-component mixture model due to the non-identifiability issues with the more general models.

There are several important implications from the mixture model. First, the mixture model differs from the usual (implicit) assumption of $(X_i^*|Y_i = 1) \sim Bin(2, \theta^*)$ with $\theta^* = \pi\theta + (1 - \pi)\theta_0$ for cases. Although $E(X) = E(X^*)$, it is shown [Zhou and Pan,

2009] that

$$E(X_i^2|Y_i=1) - E(X_i^{*2}|Y_i=1) = \pi(1-\pi)(\theta-\theta_0)^2 \geq 0,$$

where the strict inequality holds for the non-degenerated case with $\theta^* \neq \theta_0$, $\pi \neq 0$ and $\pi \neq 1$. Hence, the binomial mixture model introduces an overdispersion of the minor allele as compared to a binomial distribution with the same MAF. While the most popular 1df trend test compares the mean difference of the genotype scores between the control and case groups, it ignores the possible genetic heterogeneity in the case group and thus possible differences in high moments of the genotype scores between the two groups. Hence, taking advantage of the existing genetic heterogeneity (as modeled by the mixture model), an association test can gain power in detecting differences in both the means (i.e. first moments) and higher-order moments (e.g. second moments) between the control and case groups, which is closely related to the expanded regression model (3.2). Second, as shown by Zhou and Pan [2010], the mixture model also implies that HWE is violated in the case group under genetic heterogeneity, suggesting its connection to the HWE test.

Since complex diseases can be caused by a large number of genetic variants, it may be desirable to use a mixture model with more than two components to capture the complex heterogeneity. Qian and Shao [2013] extended the two-component mixture model to a more general form for the disease group as follows:

$$(X_i|Y_i=1) \sim \sum_{j=1}^{J} \pi_j Bin(2,\theta_j), \tag{3.6}$$

with $J \geq 2$, $0 < \theta_j < 1$ and $\pi_j \geq 0$ for any $j = 1, ..., J$, and $\sum_{j=1}^{J} \pi_j = 1$. Note that $J \geq 2$ is unknown and does not need to be specified.

To test the null hypothesis H_0: $\theta_1 = \theta_2 = \cdots = \theta_J$ (or $\pi_j = 1$ and $\pi_k = 0\ \forall k \neq j$), Qian and Shao [2013] developed a likelihood ratio test under genetic heterogeneity (LRT-H). The likelihoods L_H and L_0 are the same as before except

$$L_D = \begin{cases} \prod_{g=0}^{2} B_2(g,\hat{\theta}_1)^{n_g}, & \text{if } 4n_0n_2 \leq n_1^2; \\ \prod_{g=0}^{2} (n_g/n)^{n_g}, & \text{if } 4n_0n_2 > n_1^2. \end{cases} \tag{3.7}$$

Under H_0, the LRT-H statistic $T_{\text{LRT-H}} = 2\log(L_D L_H) - 2\log L_0$ asymptotically follows a mixture of two chi-squared distributions with 1 and 2 dfs respectively, i.e. $0.5\chi^2_1 + 0.5\chi^2_2$. Note that since the model (3.5) is not equivalent to the model (3.6), the asymptotic null distribution of the LRT-H statistics for the two models may be different.

3.3 Application to the WTCCC data

3.3.1 Quality Control

We applied the methods to the Wellcome Trust Case Control Consortium (WTCCC) data [Burton et al., 2007]. The data include two control groups, called the National UK Blood Services (NBS) and 1958 British Birth Cohort (58C). To illustrate that a conclusion is not limited to a specific disease, we considered four traits: Bipolar disorder (BD), coronary artery disease (CAD), Crohn's disease (CD), and type 2 diabetes (T2D).

We followed the quality control procedures of Burton et al. [2007] to screen for subjects and SNPs. In addition, we removed any SNP with a p-value $< 5.7 \times 10^{-7}$ by the LRT-H test contrasting the two control groups (58C vs NBS) or (NBS vs 58C); the same cutoff 5.7×10^{-7} was used by Burton et al. [2007] for the HWE test applied to the combined control group to remove SNPs. After QC, the genotyping rates were greater than 99.9% for all the four datasets (each with a combined case and control sample). The numbers of subjects in the control group (i.e. NBS+58C) before and after QC were 3,004 and 2,938 respectively. The numbers of autosomal SNPs and subjects for each disease group before and after QC are summarized as in Table 3.2.

3.3.2 GWAS Results

A genome-wide scan was conducted for each dataset with each of the four tests applied to each SNP. The corresponding Q-Q plots are shown in Figure 3.2, confirming no obvious population structures, as supported by the estimated inflation factors all close to 1 (Table 3.3).

After identifying significant SNPs at the usual genome-wide significance level of 5×10^{-8}, we applied the method of Psychiatric Genomics Consortium [PGC, 2014] to define LD-independent (index) SNPs and risk loci. Briefly, first, among all significant

SNPs, an SNP is defined to be an LD-independent SNP if it is in weak LD with $r^2 < 0.1$ with a more significant SNP within a 0.5Mb window; second, a risk locus is defined as a basepair (BP) interval including all the SNPs with $r^2 > 0.6$ to an LD-independent SNP, and any two risk loci within the distance of 0.25Mb are merged.

The numbers of significant SNPs and risk loci identified by each test are shown in Figures 3.3 and 3.4. It is clear that HWE test identified the largest numbers of significant SNPs and risk loci, most of which overlapped with those of the LRT-H test; this can be explained by the close connection between the two tests: a binomial mixture model implies the Hardy-Weinberg disequilibrium. Second, by the close relationship between the binomial mixture model and the expanded regression model (3.2), most of the significant SNPs and risk loci identified by the 2df test were also uncovered by the LRT-H test. Third, perhaps most importantly, since the LRT-H test also contrasts the allele frequency differences between the case and control groups as does the 1df trend test, the significant SNPs and risk loci identified by the popular trend test were almost all recovered by the LRT-H test. Finally, the LRT-H test also discovered some unique significant SNPs and risk loci.

Some example SNPs identified to be significant by LRT-H, but not by other tests, are shown in Table 3.4. Some significant risk loci identified by the LRT-H or HWE test, but not by the other two tests, are confirmed to be within 0.25 Mb of some previously identified SNPs or risk loci, as shown in Table 3.5. The LocusZoom plots [Pruim et al., 2010] for the significant risk loci uniquely identified by LRT-H are shown in Figures 3.5 - 3.8. To facilitate interpretation, we also added their predicted GenoCanyon scores [Lu et al., 2015]; a higher score predicts a higher likelihood for one or more SNPs in the nearby region to be functional. As GenoCanyon only supports hg19 while the original WTCCC data were all based on hg18, we lifted the annotation for the WTCCC data to hg19 using the UCSC web interface (`http://genome.ucsc.edu/cgi-bin/hgLiftOver`) before generating LocusZoom and GenoCanyon plots. We can see that many of the significant risk loci are in the regions with high functional scores.

3.4 Conclusions

We have shown possible power gain of the proposed LRT-H test which is closely related to the Hardy-Weinberg equilibrium (HWE) test. Briefly speaking, when applied to the WTCCC GWAS data, we found that the HWE test on the case group could identify the largest number of associated SNPs and risk loci for each of the four diseases considered, most of which overlapped with those of LRT-H, though LRT-H could uniquely detect some risk loci too. Although the HWE test has long been proposed as a possible choice for association testing [Nielsen et al., 1998; Wittke-Thompson et al., 2005], it has seldom been used and "often under-exploited" [Balding, 2006] for such a purpose but most widely used only for SNP genotyping quality control. It is likely that some of the significant SNPs identified by the HWE test and LRT-H are due to genotyping errors, but some may be true positives. A challenge is to validate and interpret them as for any new discovery in GWAS.

Table 3.1: The genotype frequencies for case-control data of a SNP.

	AA	Aa	aa	Total
Case	n_0	n_1	n_2	n
Control	m_0	m_1	m_2	m

Table 3.2: The QC summary. For each disease study group, both case and two control groups are included. The number of subjects for the two control groups (i.e. NBS+58C) were before and after QC were 3,004 and 2,938, respectively.

		CD	BD	CAD	T2D
Before QC	# of Subjects	2,005	1,998	1,988	1,999
After QC	# of Subjects	1,748	1,868	1,926	1,924
	# of SNPs	356,589	356,011	355,881	356,075

Table 3.3: The inflation factors of various tests.

	CD	BD	CAD	T2D
1 df GLM	1.11	1.12	1.07	1.08
2 df GLM	1.09	1.10	1.07	1.07
LRT-H	1.13	1.13	1.09	1.10
HWE	0.96	0.97	0.98	0.97

Table 3.4: Example SNPs which are identified to be significant by LRT-H, but not by HWE, 1 df Trend or 2 df Genotypic test.

Disease	SNP	CHR	BP	LRT-H	1 df Trend	2 df Genotypic	HWE
CD	rs6677092	1	238239905	1.35e-08	2.50e-04	9.51e-07	4.45e-06
BD	rs5754321	22	31600461	1.49e-08	5.31e-03	4.51e-03	2.47e-07
CAD	rs326296	3	97643917	6.28e-11	1.40e-05	2.39e-06	4.87e-07
T2D	rs9300013	11	32461118	4.62e-08	7.16e-03	2.58e-02	6.35e-07

Table 3.5: Some significant risk loci identified by HWE or LRT-H, but not by the other two tests. Each risk locus is within 250 Kb of some previously identified SNPs or loci. For each risk locus, the LD-independent SNP along with the MAF in the control group and the p-values of LRT-H ($P_{\text{LRT-H}}$) and the HWE exact test (P_{HWE}) are reported.

Disease	CHR	SNP	BP (in Mb)	$\text{MAF}_{\text{control}}$	$P_{\text{LRT-H}}$	P_{HWE}	Reported genes	Ref
CD	3	rs12714959	18.59-18.62	0.304	4.04e-05	1.51e-09	-	Franke et al (2010)
	21	rs2252931	33.61-33.63	0.266	3.07e-01	3.41e-09	IFNGR2	Jostins et al (2012)
BD	6	rs7771567	97.81-98.01	0.301	7.41e-12	1.86e-11	MIR2113/POU3F2	Mühleisen et al (2014)
T2D	3	rs2377104	185.942570	0.491	9.57e-09	2.25e-09	IGF2BP2	Hara et al (2014)
	9	rs7030479	4.202347	0.066	1.28e-12	2.26e-13	GLIS3	Mahajan et al (2014)
	11	rs1002227	17.35-17.36	0.302	3.70e-08	1.29e-08	KCNJ11	Mahajan et al (2014)
	16	rs1350889	81.342-81.345	0.478	4.79e-10	1.16e-10	CMIP	Cho et al (2012)

Table 3.6: Simulation: empirical Type I error and power when the number of independent columns (denote as "No. ind") is 25, 35, and 65. "RV.asy" and "RV.perm" standards for asympototic and permutation-based RV test, respectively.

		SPC(γ)											
No. Ind		$\gamma = 1$	2	3	4	5	6	7	8	∞	aSPC	RV.asy	RV.perm
25	Type I	0.045	0.065	0.015	0.035	0.020	0.020	0.025	0.035	0.035	0.035	0.060	0.070
	Power	0.38	0.85	0.86	0.95	0.91	0.95	0.91	0.87	0.56	0.96	0.86	0.85
45	Type I	0.040	0.050	0.045	0.030	0.035	0.035	0.035	0.040	0.060	0.020	0.055	0.055
	Power	0.16	0.52	0.57	0.76	0.72	0.76	0.65	0.69	0.37	0.71	0.52	0.52
65	Type I	0.055	0.050	0.045	0.045	0.035	0.045	0.050	0.065	0.050	0.055	0.065	0.065
	Power	0.11	0.42	0.34	0.53	0.46	0.61	0.47	0.51	0.29	0.52	0.42	0.31

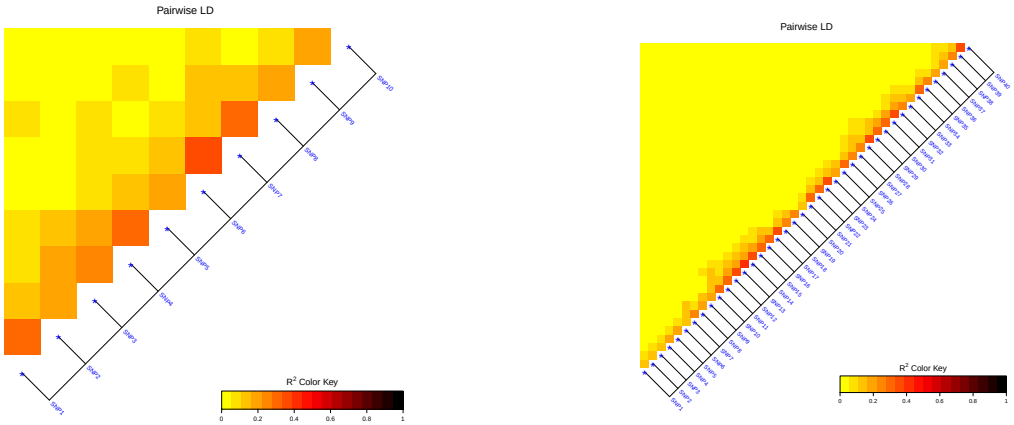


Figure 3.1: Linkage disequilibrium plots for simulated genotypes with ($n = 200$, $p = 10$) (left panel) and ($n = 400$, $p = 20$) (right panel).

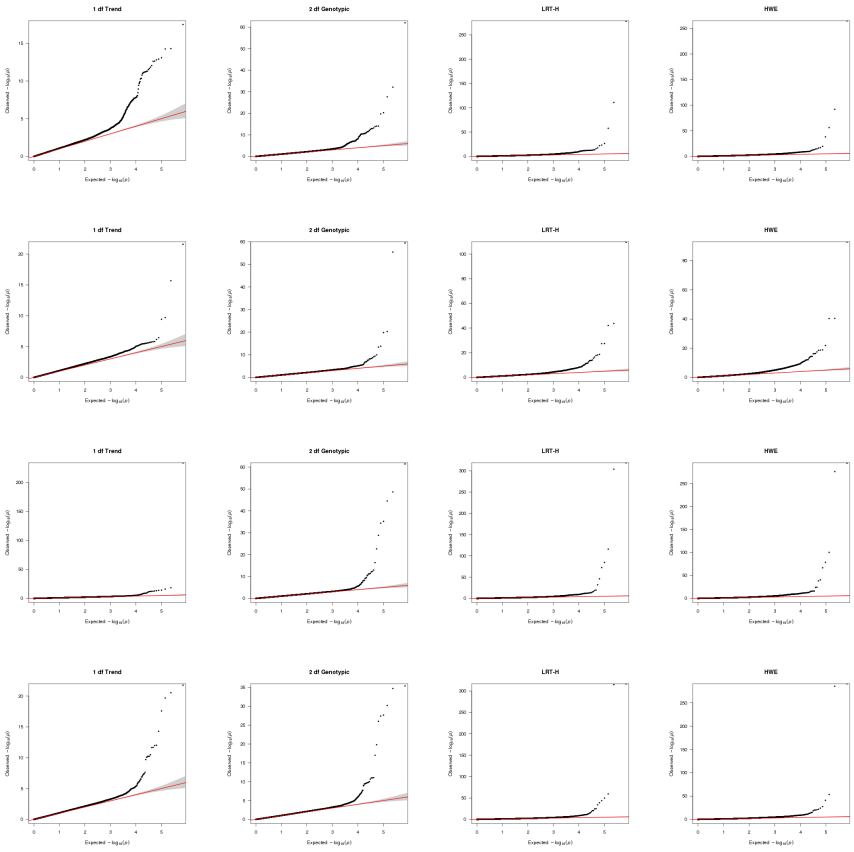


Figure 3.2: Q-Q plots of various tests for CD (first row), BD (2nd row), CAD (3rd row) and T2D (bottom row).

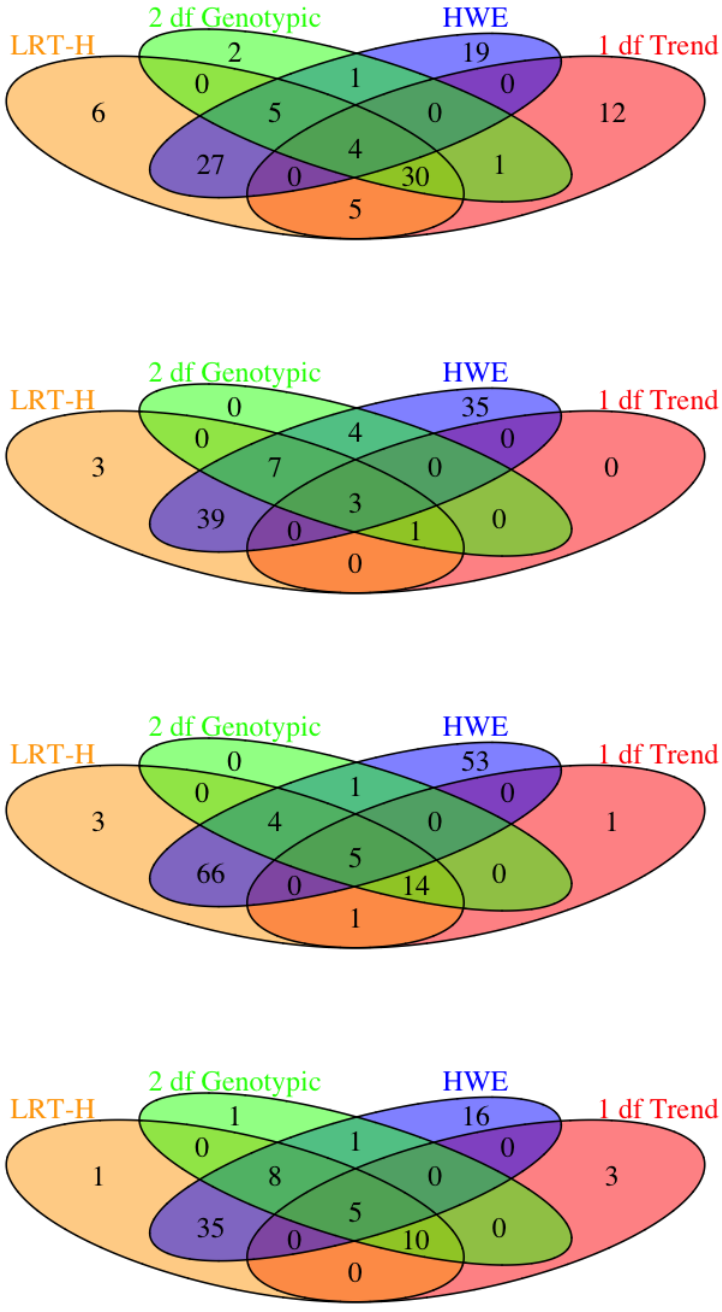


Figure 3.3: Venn-diagrams of the significant SNPs at the genome-wide significance level of 5×10^{-8} identified by each test for traits CD, BD, CAD and T2D (from the top to the bottom).

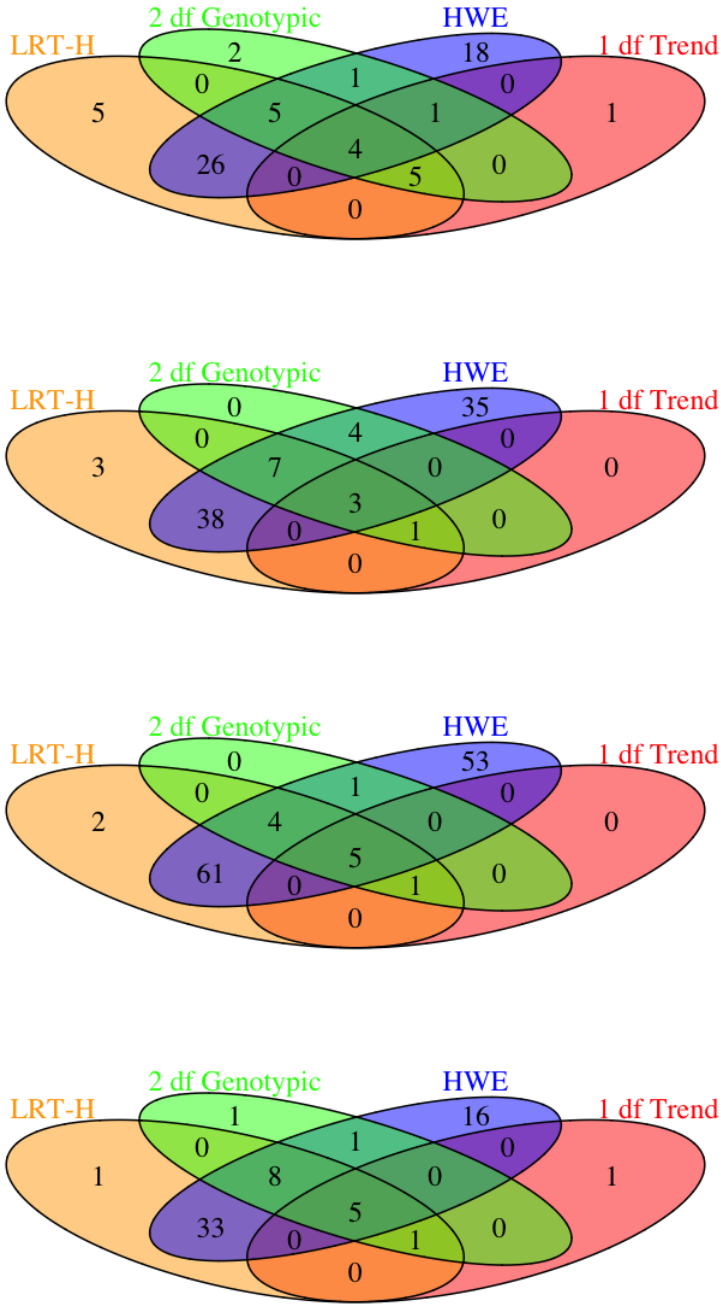


Figure 3.4: Venn-diagrams of the significant risk loci identified by each test for traits CD, BD, CAD and T2D (from the top to the bottom).

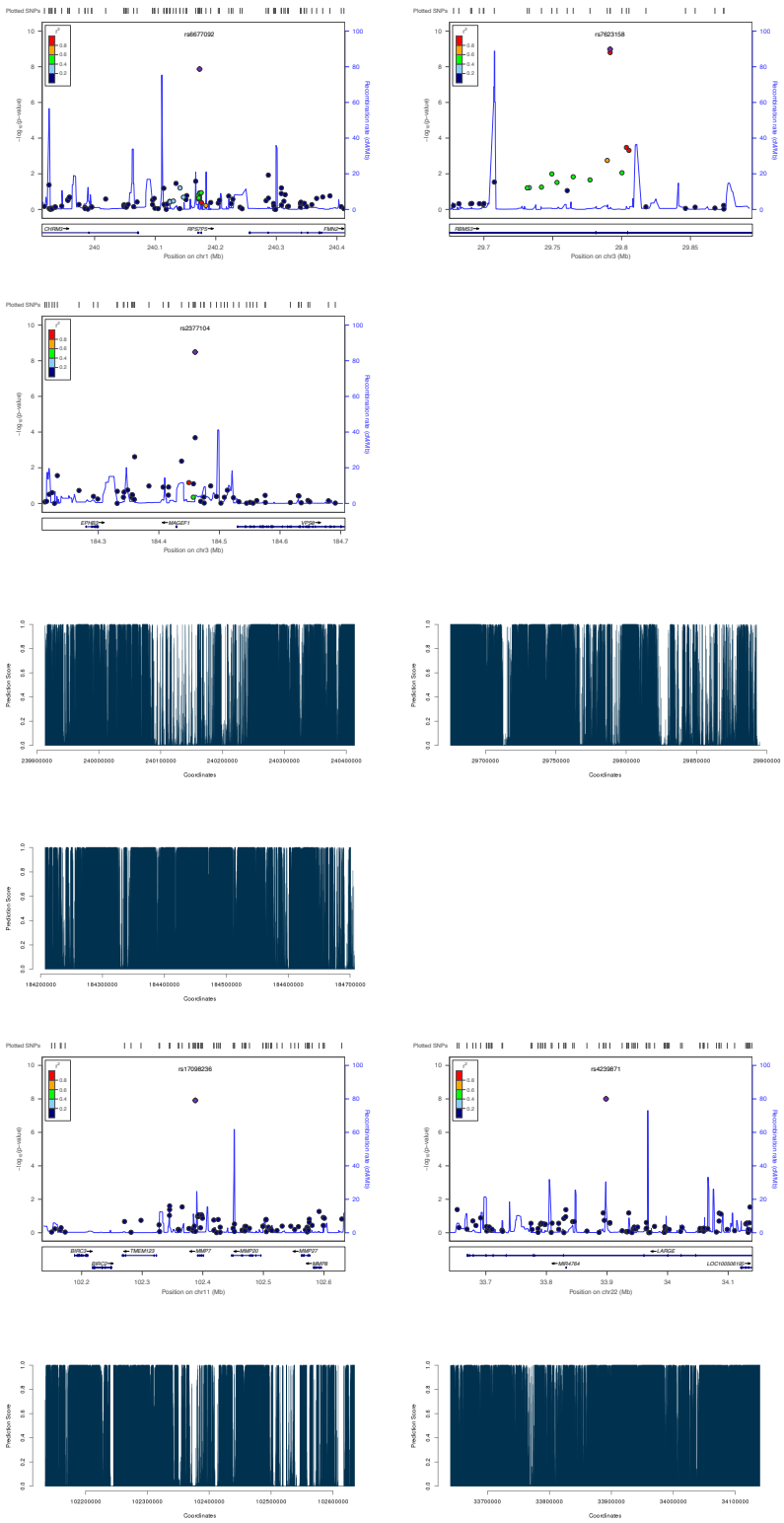


Figure 3.5: LocusZoom plots of the risk loci for trait CD, uniquely identified by LRT-H. The GenoCanyon scores for the LD-independent (index) SNPs are 2.71e-05, 5.3e-04, 1.00, 2.4e-03 and 5.5e-03 respectively.

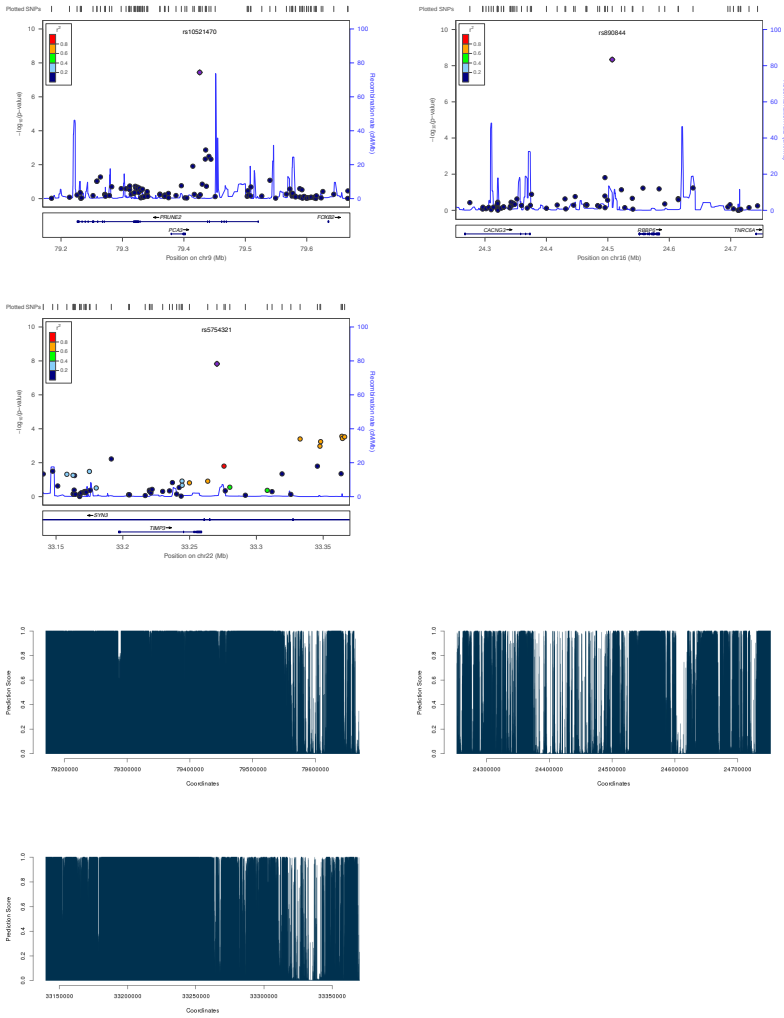


Figure 3.6: LocusZoom plots of the risk loci for trait BD, uniquely identified by LRT-H. The GenoCanyon scores for the LD-independent (index) SNPs are 0.999, 0.888 and 0.972 respectively.

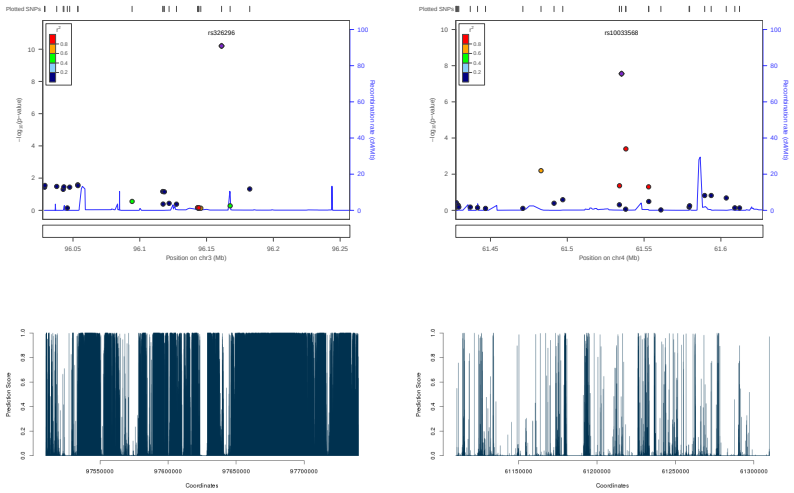


Figure 3.7: LocusZoom plots of the risk loci for trait CAD, uniquely identified by LRT-H . The GenoCanyon scores for the LD-independent (index) SNPs are 3.35e-06 and 1.24e-06 respectively.

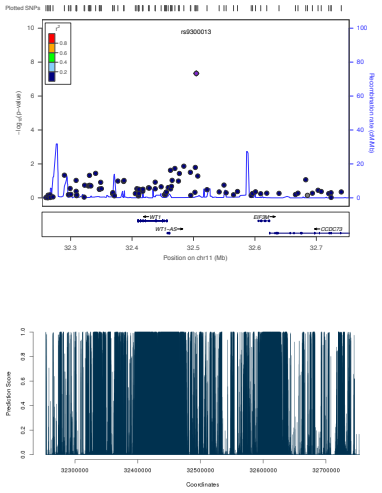


Figure 3.8: LocusZoom plot of the risk locus for trait T2D, uniquely identified by LRT-H. The GenoCanyon score for the LD-independent (index) SNP is 5.34e-05.

Chapter 4

Adaptive testing for association between two random vectors in moderate to high dimensions

4.1 Introduction

To investigate genetic control of gene expression, it is common and useful to conduct association analysis between single nucleotide polymorphisms (SNPs) and gene expression (i.e. mRNA or transcript) levels, also known as eQTL analysis. This often involves massive univariate testing. For example, Colantuoni et al. (2011) examined 30,176 expression probes and 625,439 SNPs, leading to 1.89×10^{10} (19 billion) possible SNP-gene associations. After the Bonferroni adjustment, 1,628 individual associations surpassed the genome-wide significance level. However, when they conducted a global test for possible association between all SNPs and all transcripts, no association was detected. "This dramatic lack of association between genetic distance and transcriptome distance across our sample is a surprising result that requires further interrogation. It is possible that no association is found in Fig. 4 because most of the genetic polymorphisms measured do not impact on gene expression." We agree with Colantuoni et al. (2011) on the possible reason for the lack of a global association in the presence of some individual associations: it is due to the lack of power of a global test for high-dimensional data

with only sparse signals. Furthermore, the authors also commented on that, surprisingly, no association was found even for smaller subsets of the SNPs and genes. We note that their used method was Mantel's (1967) test, which was originally proposed for low-dimensional data and may have limited powewr for moderate- to high-dimensional data as to be confirmed. Nevertheless, this example pinpoints the importance of conducting global association testing with high-dimensional data, given that most of the existing tests are almost exclusively developed for low-dimensional data for historical reasons, as reviewed in Josse and Holmes (2014).

Some commonly used tests for association between two random vectors include the RV test (Escoufier, 1970), the Mantel test (Mantel, 1967) and the dCov test (Székely et al., 2007). The RV test is based on the RV coefficient as a multivariate generalization of Pearson's correlation coefficient. It is perhaps the most popular one in many fields, especially in ecology. The Mantel test aims to detect a possible correlation between two distance matrices among the subjects based on the two random vectors respectively; it is noted that the Mantel test was used by Colantuoni et al. (2011). The dCov test has only become popular recently due to its attracting property of being consistent in detecting any possible associations, including non-linear and non-monotonic relationships. A common problem with the above tests is their treating all the variables in the two random vectors equally a priori, which is perhaps reasonable for low-dimensional data, but not for moderate- to high-dimensional data: as for the SNP-gene expression data of Colantuoni et al. (2011), most of the SNPs do not have regulatory function; even for those regulatory ones, their targets are likely only a few, not most, of the genes. That is, for high-dimensional data, we expect that many or even most (e.g. SNP-gene) pairs are not associated, which is ignored by the above existing tests, leading to their noise accumulations and thus substantial power loss as to be confirmed in later numerical studies. Hence, to boost power, it is important to conduct variable selection or variable weighting. With weak associations, it is difficult for accurate variable selection, so we take a variable weighting approach. In our approach, we use the data to adaptively determine a weight for each pair of the variables: if a pair is more likely to be associated, we assign a higher weight to it. This will effectively down-weight many of those non-associated pairs, alleviating the effects of noise accumulation hindering most existing tests for high-dimensional data. Our adaptive test can be regarded as a generalization

of the RV test to high-dimensional data, as to be shown later.

We note that the above tests aim to tackle the same problem as association testing for multiple traits or longitudinal traits in genetics (e.g., Maity et al 2012; He et al 2015; Fan et al 2016; Wang et al 2013, 2015, 2017; Kim et al 2016 and references therein), but the two lines of research seem to be largely non-overlapping; it is also our goal here to bridge the gap between the two lines of research. In particular, our proposed test is related to another adaptive test called GEE-aSPU, originally designed in genetics for testing for multi-trait and multi-SNP associations in low to moderate dimensions (Kim et al., 2016), but we will also show some computational advantages of the proposed test over GEE-aSPU. It is also connected with kernel machine regression and kernel distance methods (Hua and Ghosh, 2015). Furthermore, due to the simplicity of our proposed test, it can be also extended to detect non-linear or even non-monotonic associations by borrowing the idea from the dCov test, though our test is much more powerful than the dCov test for sparse signals in moderate- to high-dimensions.

The rest of the article is organized as follows. In section 2 we will briefly review the RV test, which serves to motivate our proposed aSPC test. We then outline the connections of the aSPC test to some existing tests before presenting its several generalizations. Section 3 applies the new and some existing tests to an SNP-gene expression data drawn from the Alzheimer's Disease Neuroimaging Initiative (ADNI), highlighting some advantages of the new tests over some existing ones. In section 4 more simulation results are shown to support the power and flexibility of the aSPC test. We end with a summary of the main conclusions in section 5.

4.2 Methods

Our goal is to test for association between two random vectors $\boldsymbol{x}_{p\times 1}$ and $\boldsymbol{y}_{q\times 1}$ in p and q dimensions respectively. We have n iid observations on $\boldsymbol{x}$-$\boldsymbol{y}$ pair as stored in two matrices $X_{n\times p}$ and $Y_{n\times q}$, respectively; each row of the two matrices corresponds to an observed $\boldsymbol{x}$-$\boldsymbol{y}$ pair. Denote $X_{\cdot l}$ as the lth ($l = 1, \ldots, p$) column of matrix X and $Y_{\cdot m}$) as the mth ($m = 1 \ldots q$) column of Y. It is assumed throughout that each column of the two matrices is centered at mean 0 with a unit variance(or Euclidean norm). We will use X and Y to test for association between $\boldsymbol{x}$ and $\boldsymbol{y}$; with some abuse of notation, we

also call association between X and Y.

4.2.1 Review: the RV test

The two cross-product matrices of X and Y are $W_X = XX^T$ and $W_Y = YY^T$, both of which are of size $n \times n$. To measure their proximity, the Hilbert-Schmidt inner product between matrices W_X and W_Y can be used:

$$< W_X, W_Y >= tr(XX^TYY^T) = \sum_{l=1}^{p}\sum_{m=1}^{q} \mathrm{Cov}_n^2(X_{\cdot l}, Y_{\cdot m}), \tag{4.1}$$

where $\mathrm{Cov}_n(X_{\cdot l}, Y_{\cdot m})$ is the sample covariance between columns $X_{\cdot l}$ and $Y_{\cdot m}$. The RV coefficient, a correlation coefficient proposed by Escoufier (1973) for two random vectors, is computed by normalizing the Hilbert-Schmidt inner product by the matrix norms:

$$\mathrm{RV}(X, Y) = \frac{< W_X, W_Y >}{||W_X||||W_Y||} = \frac{\mathrm{tr}(XX^TYY^T)}{\sqrt{\mathrm{tr}(XX^T)^2\mathrm{tr}(YY^T)^2}}, \tag{4.2}$$

which accounts for possibly different scales of $\boldsymbol{x}$ and $\boldsymbol{y}$. The population RV coefficient is $\rho(\boldsymbol{x}, \boldsymbol{y}) = \mathrm{tr}(\Sigma_{\boldsymbol{xy}}\Sigma_{\boldsymbol{yx}})/\sqrt{\mathrm{tr}(\Sigma_{\boldsymbol{xx}}^2)\mathrm{tr}(\Sigma_{\boldsymbol{yy}}^2)}$, where $\Sigma_{\boldsymbol{xy}}$ is the population covariance between $\boldsymbol{x}$ and $\boldsymbol{y}$. Our goal is to test $H_0 : \rho(\boldsymbol{x}, \boldsymbol{y}) = 0$.

If each column of X and of Y is standardized to have a zero mean and a unit variance, as always assumed here, the RV coefficient can be simplified as:

$$\mathrm{RV}(X, Y) = \frac{\mathrm{tr}(XX^TYY^T)}{pq} = \frac{\sum_{l=1}^{p}\sum_{m=1}^{q} \mathrm{corr}_n^2(X_{\cdot l}, Y_{\cdot m})}{pq} \propto \sum_{l=1}^{p}\sum_{m=1}^{q} \mathrm{corr}_n^2(X_{\cdot l}, Y_{\cdot m}), \tag{4.3}$$

where $\mathrm{corr}_n(X_{\cdot l}, Y_{\cdot m})$ is the sample Pearson correlation coefficient between columns $X_{\cdot l}$ and $Y_{\cdot m}$.

A permutation method can be used to calculate the P-value. Specifically, for each permutation $b = 1, \ldots, B$, we permute the rows of matrix X (or Y), then calculate the corresponding RV coefficient $\mathrm{RV}^{(b)}$; the P-value is calculated as the sample proportion $(\sum_{n=1}^{B} I(RV \leq RV^{(b)}) + 1)/(B + 1)$.

4.2.2 New method: an adaptive sum of powered correlation (aSPC) test

To generalize the RV coefficient as reformulated in equation (4.3), we propose a family of so-called sum of powered correlation (SPC) tests:

$$\text{SPC}(\gamma) = \sum_{l=1}^{p}\sum_{m=1}^{q} \text{corr}_n^{\gamma}(X_{\cdot l}, Y_{\cdot m}) \tag{4.4}$$

for a set of integers $\gamma \geq 1$. Each term $\text{corr}_n^{\gamma}(X_{\cdot l}, Y_{\cdot m})$ in equation (4.4) can be re-written as $\text{corr}_n^{\gamma}(X_{\cdot l}, Y_{\cdot m}) = w_{lm}\text{corr}_n(X_{\cdot l}, Y_{\cdot m})$, where $w_{lm} = \text{corr}_n^{\gamma-1}(X_{\cdot l}, Y_{\cdot m})$ is regarded as a weight for $\text{corr}_n(X_{\cdot l}, Y_{\cdot m})$. Therefore, a larger $|\text{corr}_n(X_{\cdot l}, Y_{\cdot m})|$ will a yield higher weight $|w_{lm}|$, which will help improve power with sparse alternatives that are common for moderate- to high-dimensional data. Specifically, when $\gamma = 1$, all $\text{corr}_n(X_{\cdot l}, Y_{\cdot m})$'s will be assigned an equal weight 1, which will be beneficial for dense alternatives (i.e. if all or most of the columns of the two matrices X and Y are associated); however, when $\gamma \geq 2$, the larger the γ, the higher weights would be assigned to those larger $\text{corr}_n(X_{\cdot l}, Y_{\cdot m})$'s, more and more favoring sparse alternatives (i.e. when only few of the columns of X and Y, as indicated by those larger $\text{corr}_n(X_{\cdot l}, Y_{\cdot m})$'s, are truly associated with each other). In the extreme case of a sparse alternative with only one or few associated column-pairs between X and Y, for an even integer $\gamma \to \infty$, we have

$$\text{SPC}(\gamma) \propto \left(\sum_{l=1}^{p}\sum_{m=1}^{q} \text{corr}_n^{\gamma}(X_{\cdot l}, Y_{\cdot m})\right)^{1/\gamma} \to \max_j |\text{corr}_n(X_{\cdot l}, Y_{\cdot m})| = \text{SPC}(\infty), \tag{4.5}$$

which we can see largely eliminates the effects of non-associated pairs and thus is expected to be more powerful for more sparse alternatives. We emphasize that, with large p and q in moderate to high dimensions, noise accumulation is a severe problem for sparse alternatives, which explains power loss of many non-adaptive tests like the RV test, as to be shown later.

In summary, depending on the type of a true alternative hypothesis to be tested, i.e. dense or sparse, a small or a large γ would yield higher power for the SPU(γ) test. In practice, because it is unknown what is the true alternative and thus which γ value would yield high power, we develop an adaptive SPC (aSPC) test to combine the

evidence across the SPC tests:

$$\text{aSPC} = \min_{\gamma \in \Gamma} P_{\text{SPC}(\gamma)} \tag{4.6}$$

where $P_{\text{SPC}(\gamma)}$ is the P-value of the SPC(γ) test, and Γ contains a set of candidate values for γ. In general, $\Gamma = \{1, 2, \ldots, \gamma_u, \infty\}$ with $1 < \gamma_u < \infty$ can be used; larger p and q require a larger γ_u; a practical guideline on the choice of γ_u is that SPC(γ_u) gives results similar to SPC(∞). We used $\Gamma = \{1, \ldots, 8, \infty\}$ throughout this paper for its good performance based on our limited experience.

A permutation method can be used to obtain the P-values of all the SPC and aSPC tests in a *single loop* (or layer) of permutations. Briefly, B copies of the null statistic $\text{SPC}(\gamma)^{(b)}$ for each $\gamma \in \Gamma$ and $b = 1, \ldots, B$ can be calculated by permuting the rows of matrices X (or Y) B times. The P-value of each SPC(γ) is calculated as $P_{\text{SPC}(\gamma)} = (\sum_{b=1}^{B} I(|\text{SPC}(\gamma)^{(b)}| \geq |\text{SPC}(\gamma)|) + 1)/B$. Furthermore, based on the same B copies of the null statistics, we calculate the P-value for the aSPU test as $P_{\text{aSPC}} = (\sum_{b=1}^{B} I(\text{aSPC}^{(b)} \leq \text{aSPC}) + 1)/(B + 1)$ with $\text{aSPC}^{(b)} = \min_{\gamma \in \Gamma} p_{\gamma}^{(b)}$ and $p_{\gamma}^{(b_1)} = (\sum_{b \neq b_1} I(|\text{SPC}(\gamma)^{(b)}| \geq |\text{SPC}(\gamma)|)^{(b_1)} + 1)/B$.

4.2.3 Connections with some existing tests

We start by establishing a relationship between the aSPC test (with the Pearson correlation coefficient) and an existing test called GEE-aSPU, which was proposed by Kim et al. (2016) for multiple trait-multiple SNP associations. We first review the GEE-aSPU test before pointing out its connection to the aSPC test.

First we need some notations. Denote $X_{i\cdot} = (x_{i1}, \ldots, x_{ip})$ and $Y_{i\cdot} = (y_{i1}, \ldots, y_{iq})^T$ as ith row in matrices X and transpose of ith row in Y for $i = 1, \ldots, n$, respectively; denote $X_i = I \otimes X_{i\cdot}$, where I is a $q \times q$ identity matrix, and $\otimes$ represents the Kronecker product.

Suppose we treat each column $Y_{\cdot m}$ for $m = 1, \ldots, q$ in $Y_{n \times q}$ as a response, each column $X_{\cdot l}$ for $l = 1, \ldots, p$ in $X_{n \times p}$ as a covariate or predictor of interest; recall that $Y_{\cdot m}$ and $X_{\cdot l}$ has been standardized to have zero mean and unit variance. We can then test if there is any association between the columns of X and those of Y with a marginal

generalized linear model

$$g(E(Y_{i\cdot}|X_i)) = X_i\beta, \tag{4.7}$$

where $g(.)$ is a canonical link function, and β is a pq-dimensional vector of unknown parameters of interest. We aim to test the null hypothesis $H_0 : \beta = 0$. Denote $\bar{Y}$ as the mean vector of columns of Y, which is a zero vector of length q. With a canonical link function and a working independence model in GEE (Liang and Zeger, 1986), the generalized score vector for β is

$$U = \sum_{i=1}^{n} X_i^T(Y_{i\cdot} - \bar{Y}) = \sum_{i=1}^{n} X_i^T Y_{i\cdot}. \tag{4.8}$$

It is easy to verify $U = (U_{11}, \ldots, U_{p1}, \ldots, U_{1q}, \ldots, U_{pq})^T$ with $U_{lm} = X_{\cdot l}^T Y_{\cdot m} = \mathrm{corr}_n(X_{\cdot l}, Y_{\cdot m})$. That is, each element U_{lm} measures the association between columns $X_{\cdot l}$ and $Y_{\cdot m}$. The GEE-SPU test statistic is defined by

$$\mathrm{SPU}(\gamma_1, \gamma_2) = \sum_{m=1}^{q} \left[\left(\sum_{l=1}^{p} U_{lm}^{\gamma_1} \right)^{\frac{1}{\gamma_1}} \right]^{\gamma_2} = \sum_{m=1}^{q} \left[\left(\sum_{l=1}^{p} \mathrm{corr}_n^{\gamma_1}(X_{\cdot l}, Y_{\cdot m}) \right)^{\frac{1}{\gamma_1}} \right]^{\gamma_2}. \tag{4.9}$$

Denote Γ_1 and Γ_2 are two sets of positive integers. The GEE-aSPU test statistic is then defined as the minimum p-value of SPU(γ_1, γ_2)' tests for all $\gamma_1 \in \Gamma_1$ and $\gamma_2 \in \Gamma_2$:

$$\mathrm{aSPU} = \min_{\gamma_1, \gamma_2} p_{\gamma_1, \gamma_2} \tag{4.10}$$

Here we observe a close connection between the SPC test and the GEE-SPU test: if $\gamma_1 = \gamma_2 = \gamma$, we have $\mathrm{SPU}(\gamma, \gamma) = \mathrm{SPC}(\gamma)$. The difference between aSPC and aSPU tests is that the latter searches for two optimal (γ_1, γ_2) in a two-dimensional space (i.e. over $\Gamma_1 \times \Gamma_2$), while aSPC searches over only a one-dimensional space (i.e. Γ); the GEE-aSPU test reduces to aSPC if we impose $\gamma_1 = \gamma_2 = \gamma$.

Due to the currently inefficient implementation of the GEE-aSPU test (for its general regression framework) in R package `GEE-aSPU`, it cannot be applied to high-dimensional data: it requires too much memory space for its inefficient storage of the design matrix with dimension $np \times pq$ (or $nq \times pq$) if Y (or X) is treated as the response. As an example, the GEE-aSPU test will need about 40Gb of memory if $p = q = 300$ and

sample size $n = 200$, not yet affordable by many computers. In contrast, due to its simplicity, the aSPC test is applicable to high-dimensional data.

Finally, we comment on that the SPC(2) test is also closely related to several other tests, further illustrating the potential power of the aSPC test. First, since the dCov test and the Hilbert-Schmidt independence criterion (HSIC) test are equivalent (Sejdinovic et al., 2013), Hua and Ghosh (2015) called them kernel distance covariance method (KDC); they further established the equivalence of KDC and multivariate kernel machine regression (KMR) test (Maity et al., 2012) (if the same kernels are used in the two). On the other hand, Kim et al. (2016) pointed out that GEE-SPU(2,2) is similar to multivariate KMR with a linear kernel; the two are exactly the same if the true correlation matrix is used as the working correlation structure in GEE for the former, which in general does not hold (unless the columns of Y are independent), because the working independence model is used in GEE-SPU tests. Now, by the equivalence between SPC(2) and GEE-aSPU(2,2) and by the above results, we see the close similarity between SPC(2) and other tests. Using the weighting argument motivating the development of other SPC(γ) tests with $\gamma > 2$, we expect that the other tests (i.e. dCov, HSIC and KMR with linear kernels) may lose power with sparse association patterns, which is to be confirmed in our later simulations.

4.2.4 Extensions

So far we define an SPC test with the Pearson correlation coefficients between the columns of the two matrices. Here we generalize the SPC and thus aSPC tests with several other dependence measures and with covariates.

Fisher's transformation

We may take Fisher's z-transformation on the sample Pearson correlation coefficient $r_{lm} = \text{corr}_n(X_{\cdot l}, Y_{\cdot m})$ before plugging into equation (4.4). The reason is to account for heterogeneous variances of the sample correlations for an alternative hypothesis; as to be shown next, the variance of a sample correlation increases monotonically as the absolute value of the true correlation decreases (under the normality assumption). Specifically, the sample correlation $r_{lm} = \text{corr}_n(X_{\cdot l}, Y_{\cdot m})$ is replaced by $z_{lm} = \frac{1}{2}\ln\left((1 + r_{lm})/(1 - r_{lm})\right)$ in equation (4.4). Under the normality assumption (on each

pair of the columns of X and Y), z_{lm} is approximately normally distributed with mean $\frac{1}{2}\ln((1+\rho_{lm})/(1-\rho_{lm}))$ and a constant variance $1/(n-3)$, where ρ_{lm} is the population Pearson correlation coefficient.

Given that $z_{lm} \dot{\sim} N(\frac{1}{2}\ln((1+\rho_{lm})/(1-\rho_{lm})), 1/(n-3))$, it is not hard to find the approximate distribution of the sample Pearson correlation coefficient is $r_{lm} \dot{\sim} N(\rho_{lm}, (1-\rho_{lm}^2)^2/(n-3))$; the variance $(1-\rho_{lm}^2)^2/(n-3)$ is obtained by the delta method and clearly confirms the monotonicity mentioned above. In particular, since the variance is largest for no correlations, not taking Fisher's transformation or not stabilizing the variance may lead to loss of power, especially for high-dimensional data, for which sparse alternatives are expected with many non-associated pairs.

Whenever needed, to distinguish using Fisher's z-transformed Pearson correlation coefficients from using other dependence measures for the SPC and aSPC tests, we will use SPC.P and aSPC.P to refer to the former:

$$\text{SPC.P}(\gamma) = \sum_{l=1}^{p}\sum_{m=1}^{q} z_{lm}^{\gamma}, \tag{4.11}$$

and the aSPC.P test is similarly defined as before.

The aSPC test with Spearman's correlation

More generally, the sample Pearson correlation coefficient term $r_{lm} = \text{corr}_n(X_{\cdot l}, Y_{\cdot m})$ in equation (4.4) can be replaced by a different dependence measure. For example, we can use Spearman's (1904) rank correlation coefficient, which is effective for monotonic relationships, in contrast to only linear relationships by Pearson's coefficient. The Spearman correlation coefficient is defined as the Pearson correlation coefficient between the ranked variables. Specifically, $X_{\cdot l}$ and $Y_{\cdot m}$ ($l = 1, \ldots, p$ and $m = 1, \ldots, q$) are converted to the rank score vectors $\text{rank}(X_{\cdot l})$ and $\text{rank}(Y_{\cdot m})$ (e.g. rank score = 1 for the smallest value in $X_{\cdot l}$ (or $Y_{\cdot m}$) and rank score = n for the largest value in $X_{\cdot l}$ or $(Y_{\cdot m})$). The sample Spearman correlation coefficient is calculated as

$$r_{lm}(\text{Spearman}) = \frac{\text{Cov}_n(\text{rank}(X_{\cdot l}), \text{rank}(Y_{\cdot m}))}{\sqrt{\text{Cov}_n(\text{rank}(X_{\cdot l}), \text{rank}(X_{\cdot l}))\text{Cov}_n(\text{rank}(Y_{\cdot m}), \text{rank}(Y_{\cdot m}))}}, \tag{4.12}$$

where $\text{Cov}_n(u, v)$ is a sample covariance between vectors $u_{n\times 1}$ and $v_{n\times 1}$. Then the SPC statistic with Spearman's rank correlation coefficient is defined as:

$$T_{\text{SPC.Sp}(\gamma)} = \sum_{l=1}^{p}\sum_{m=1}^{q} r_{lm}^{\gamma}(\text{Spearman}), \tag{4.13}$$

and aSPC.Sp is defined similarly as before.

The aSPC test with the distance correlation

Another extension is to replace each sample Pearson correlation coefficient in equation (4.4) by a corresponding distance correlation coefficient (dCor), which is derived based on the distance covariance (dCov) (Szykely et al., 2007) and is consistent in detecting any dependency, not only the linear ones (detectable by Pearson's) or monotonic ones (by Spearman's); for example, in the presence of non-linear (and non-monotonic) dependency, use of dCor is expected to be more powerful, as to be confirmed in our later simulations. We first review the usual dCov test and then modify the SPC test with the distance correlations.

The standard dCov test utilizes all columns in X and Y to calculate the pairwise distance before computing the sample distance covariance:

$$a_{ij} = ||X_{i\cdot} - X_{j\cdot}||^t, b_{ij} = ||Y_{i\cdot} - Y_{j\cdot}||^t, \tag{4.14}$$

where $||\cdot||$ denotes the Euclidean distance/norm; $X_{i\cdot}$ and $Y_{i\cdot}$ denote the ith row of X and Y respectively ($i = 1, \ldots, n$); $t \in (0, 2]$ and $t = 1$ corresponds to the Euclidean norm, which was used in our data analysis throughout unless specified otherwise. The pairwise distances are doubly centered:

$$A_{ij} = a_{ij} - \bar{a}_{i\cdot} - \bar{a}_{\cdot j} + \bar{a}_{\cdot\cdot}, B_{ij} = b_{ij} - \bar{b}_{i\cdot} - \bar{b}_{\cdot j} + \bar{b}_{\cdot\cdot}, \tag{4.15}$$

where $\bar{a}_{i\cdot}$, $\bar{a}_{\cdot j}$ and $\bar{a}_{\cdot\cdot}$ are the ith row mean, the jth column mean and the grand mean of matrix $[a_{ij}]$; $\bar{b}_{i\cdot}$, $\bar{b}_{\cdot j}$ and $\bar{b}_{\cdot\cdot}$ are similar defined for matrix $[b_{ij}]$. Then the squared sample

distance covariance of X and Y is defined as:

$$\mathrm{dCov}_n^2(X,Y) = \frac{1}{n^2}\sum_{i,j=1}^{n} A_{ij}B_{ij}. \tag{4.16}$$

A permutation method can be used to calculate the P-value. The null statistics $T_{\mathrm{dCov}}^{(b)} = \frac{1}{n^2}\sum_{i,j=1}^{n} A_{ij}^{(b)}B_{ij}^{(b)}$ can be calculated based on each permuted sample $X^{(b)}$ and $Y^{(b)}$, where $X^{(b)}$ (or $Y^{(b)}$) is generated by permuting the rows of X (or Y). The P-value is calculated as $P_{\mathrm{dCov}} = (\sum_b^B I(\mathrm{dCov}^{(b)} \geq \mathrm{dCov}) + 1)/(B+1)$ based on B permutations.

In the standard dCov test, all columns of X and Y are used to calculate the pairwise distances; that is, each variable (or dimension) is treated equally a priori, which may not be a good idea for high-dimensional data for the abundance of sparse alternatives. In contrast, in our SPC test, each column/variable of X and Y is treated differently according to the magnitudes of their estimated pairwise associations. Specifically, similar to the standard dCov test, first we define all pairwise distances among the observations based on the ith and jth elements of $X_{\cdot l}$ and $Y_{\cdot m}$ as

$$a_{ij(l)} = ||X_{il} - X_{jl}||^t, b_{ij(m)} = ||Y_{im} - Y_{jm}||^t, \tag{4.17}$$

which computes the $n \times n$ distance matrices $(a_{ij(l)})$ and $(b_{ij(m)})$ for $i = 1, \ldots, n$, $j = 1, \ldots, n$, $l = 1, \ldots, p$ and $m = 1, \ldots, q$. Denote $\bar{a}_{i\cdot(l)}$, $\bar{a}_{\cdot j(l)}$ and $\bar{a}_{\cdot\cdot(l)}$ as the ith row mean, the jth column mean and the grand mean of $[a_{ij(l)}]$; similarly, denote $\bar{b}_{i\cdot(m)}$, $\bar{b}_{\cdot j(m)}$ and $\bar{b}_{\cdot\cdot(m)}$ for $[b_{ij(m)}]$. The elements $a_{ij(l)}$ and $b_{ij(m)}$ are then doubly centered as:

$$A_{ij(l)} = a_{ij(l)} - \bar{a}_{i\cdot(l)} - \bar{a}_{\cdot j(l)} + \bar{a}_{\cdot\cdot(l)}, B_{ij(m)} = b_{ij(m)} - \bar{b}_{i\cdot(m)} - \bar{b}_{\cdot j(m)} + \bar{b}_{\cdot\cdot(m)}, \tag{4.18}$$

then the squared sample distance covariance is defined as:

$$\mathrm{dCov}_n^2(X_{\cdot l}, Y_{\cdot m}) = \frac{1}{n^2}\sum_{i,j=1}^{n} A_{ij(l)}B_{ij(m)}. \tag{4.19}$$

The sample distance correlation (dCor) between $X_{\cdot l}$ and $Y_{\cdot m}$ is then defined as

$$\mathrm{dCor}_n(X_{\cdot l}, Y_{\cdot m}) = \frac{\mathrm{dCov}_n(X_{\cdot l}, Y_{\cdot m})}{\sqrt{\mathrm{dCov}_n(X_{\cdot l}, X_{\cdot l})\mathrm{dCov}_n(Y_{\cdot m}, Y_{\cdot m})}}. \tag{4.20}$$

The SPC.dCor test statistic is defined as:

$$\text{SPC.dCor}(\gamma) = \sum_{l=1}^{p} \sum_{m=1}^{q} \text{dCor}_n^{\gamma}(X_{\cdot l}, Y_{\cdot m}) \tag{4.21}$$

and the aSPC.dCor is similarly defined as before.

As to be shown later in simulations, the aSPC.dCor test was much more powerful than the standard distance covariance (dCov) test for sparse alternatives in even only moderate dimensions, presumably because the former's weighting on the pairwise dCor's alleviates the harmful effects of noise accumulations in the latter.

The aSPC test with covariates

The aSPC test can be applied to situations with covariates. We only need to first regress X and/or Y on the covariates, then use the residuals to construct the SPC tests. We will illustrate such an application in the example section.

4.2.5 Software

The asymptotic- and permutation-based RV tests are available as functions `coeffRV()` and `RV.rtest()` in R packages `FactoMineR` and `ade4`, respectively. The permutation-based Mantel test, dCov test and GEE-aSPU test are in functions `mantel()`, `dcov.test()`, `GEEaSPUset()` in R packages `vegan`, `energy` and `GEEaSPU`, respectively. We implemented various versions of the new SPC and aSPC tests in an R package `aSPC`, which is available on github (and CRAN).

4.3 Real data application

4.3.1 Testing for SNP-gene expression associations

To understand gene regulation, it is important to detect genetic variants like Single Nucleotide Polymorphisms (SNPs) that are associated with gene expression (i.e. transcript) levels, called eQTL (Minas et al., 2013). Due to the relatively small sample size and a severe penalty on multiple testing for a large number of SNP-gene pairs, it is often

low-powered to detect many associations at the individual apir level. As an alternative, we may first test the association between a set of SNPs and a set of the genes.

The ADNI genotype data consist of 757 subjects from ADNI-1, two hundred and thirty six of whom also have genome-wide gene expression data based on the whole blood. A pathway for Alzheimer's disease (hsa05010, `http://www.genome.jp/dbget-bin/www_bget?hsa05010`) was downloaded from Kyoto Encyclopedia of Genes and Genomes (KEGG) website (Kanehisa et al., 2016). Since the ADNI-1 genotype data are based on the human genome version hg18, we used the hg18 gene coordinate file downloaded from the PLINK website (`http://pngu.mgh.harvard.edu/~purcell/plink/`) to identify the starting base pair (bp) and ending bp for each gene. We then extracted two sets of the SNPs for the genes in the AD pathway. In the first, the SNPs within each gene were selected, including possibly both protein coding and regulatory SNPs; in the second, to focus on only regulatory SNPs, only the SNPs within the upstream 20kb of a gene's starting bp or within the downstream 20kb of its ending bp were selected. Since the results were similar, we will discuss only the first dataset.

To account for possible effects of age and gender on gene expression, we used a linear regression model to regress each gene's expression level on the two covariates, then used the residuals as the gene's adjusted expression levels in the subsequent analysis. In the end, there were 441 probes corresponding to 151 genes, and 2,483 SNPs (after excluding those with a minor allele frequency less than 0.05) in the first dataset.

To demonstrate the effects of association patterns, especially the signal sparsity levels, on the testing results, we screened the SNP-gene pairs using each pair's P-value for their marginal association, which was based on a simple linear regression of each gene's adjusted expression level on each SNP in the set. The expression level of each gene was calculated as the average of its corresponding probes for those genes with more than one probe. We used various threshold values to select subsets of the SNP-gene pairs, with a marginal P-value smaller than a given threshold. Then we pooled the SNPs and the probes in the genes surviving such a screening into a SNP set and a probe set respectively, then tested their associations using various methods. For any permutation-based test, we used a permutation number $B = 1 \times 10^4$ (unless specified otherwise). As the dimensions of the probes and the SNPs were high (i.e. in hundreds to thousands), it would be infeasible to run the GEE-aSPU test as it required a too

large memory space. The results are summarized in Table 4.1.

We have the following observations. First, when we included all the SNPs and the probes (with a P-value threshold 1), the aSPC tests (i.e. aSPC.P, aSPC.Sp, and aSPC.dCor) all gave significant P-values; in contrast, none of the other tests, including the RV test, the Mantel test and dCov test, gave any significant P-value less than the nominal level 0.05. Second, most strikingly, regardless of the dimensions (p, q) with various threshold values, the aSPC tests consistently gave small and significant P-values (e.g. < 0.001), showing their robustness to the varying association patterns (e.g. signal sparsity levels); in contrast, as fewer and fewer, but more significant, SNPs and probes were included, other global tests gradually gave more and more significant P-values, suggesting their loss of power in the presence of sparse signals due to their none-adaptiveness. Third, among the SPC tests, those SPC.P(γ) tests with larger γ (e.g. $\gamma >= 4$) gave more significant P-values than those with smaller γ (e.g. $\gamma < 4$), indicating sparse signals as expected (i.e. most SNP-probe pairs were not associated).

4.4 Simulations

4.4.1 Simulation I: linear associations

To further investigate the operating characteristics of the proposed tests, we compare their power performance with several existing tests. We first consider the ideal situation with a linear association between two sets of normal variates.

To generate a simulated dataset, two matrices $X_{n\times p}$ and $Y_{n\times p}$ were simulated with $n = 500$. First, for each X and Y, p (= 25 ,45 or 65) independent columns were simulated from a standard multivariate normal distribution. Second, a matrix $Z_{n\times 10}$ with ten columns were simulated from a multivariate normal distribution with mean 0 and a compound symmetry covariance matrix (with all diagonal elements equal to 1 and all off-diagonal elements equal to 0.1); for power comparisons, we added the first 5 columns of Z to X and the last 5 columns of Z to Y.

We applied the aSPC.P, aSPC.Sp, aSPC.dCor, RV, Mantel and dCov tests to each simulated dataset, and and compared their empirical Type I error and power estimates. The Mantel and dCov tests were conducted with the Euclidean distance. We set $B = 1000$ for any permutation-based tests. To save computing time, the empirical Type I

error rates and power of aSPC.dCor were based on 1,000 replicates while for all other tests, they were based on 10,000 replicates.

As shown in Table 4.2, first, the Type I error rates were in general well controlled for each test. Second, among all the tests, GEE-aSPU was most powerful, followed by aSPC.P. Note that, due to the linear association, aSPC.P is expected to be more powerful than aSPC.Sp (and aSPC.dCor). Third, SPC.P(2) gave the results essentially the same as both the asymptotic and permutation-based RV tests, as expected. Fourth, due to the presence many independent columns in the two matrices X and Y, a SPC.P test with a larger and finite γ (e.g. $\gamma = 6$) was more powerful than that with a small $\gamma \leq 4$; their power difference increased with the number of independent columns. Fifth, aSPC.dCor gave much higher power than dCov test, due to that SPC.dCor(γ) with larger γ reduced the effects of noise accumulation with independent columns. Moreover, we note the extremely low power of the Mantel test, followed by MANOVA.

To assess the computing time and feasibility for the permutation-based RV, GEE-aSPU and aSPC tests, we changed the number of columns in X and Y to 30, 50, 70 and 100 respectively, and with a sample size $n = 200$. We then calculated the computing time with a permutation number $B = 1 \times 10^3$. Note that, for example, for $p = q = 300$, GEE-aSPU needs to construct a large design matrix with dimension $60,000 \times 90,000$, requiring about 40GB of memory. The computing time was based on one processor (Intel Haswell E5-2680v3 with 2.5GB of memory on Unix system) from a cluster at the Minnesota Supercomputing Institute (MSI).

As shown in Figure 4.1, first, our implementation of aSPC.P completely in R was even faster than the RV.perm test, which was surprising given that aSPC.P involved conducting SPC.P(γ) for $\gamma = 1, \ldots, 8, \infty$ and RV.perm is equivalent to SPC.P(2). Second, aSPC.dCor was more computing-intensive than other tests; for data matrices $X_{n\times p}$ and $Y_{n\times q}$, aSPC.dCor required calculating pairwise distance covariances pq times based on $p + q$ distance matrices, even if we used more memory space to save the distance matrices in our current implementation in R.

4.4.2 Simulation II: non-linear associations

Now we consider a more challenging case with a non-linear and non-monotonic association. Our simulation set-up was similar to that of Székely et al. (2007).

Data matrix $X_{n\times 5}$ was simulated from a multivariate standard normal distribution. To calculate the empirical type I error rates, for each replicate a matrix $Y_{n\times 5}$ was simulated from a multivariate standard normal distribution. For power, $Y_{n\times p}$ was generated such that each of the first p_0 ($p_0 = 1, 2, 3, 4$ or 5) columns $Y_{ij} = log(X_{ij}^2)$ for $j = 1, \ldots, p_0$ and $i = 1 \ldots n$; when $p_0 \leq 4$, each of the other columns of $Y_{n\times p}$ was independently and identically simulated from a standard normal distribution. We were interested in how the empirical power changed as the number of non-linearly associated column pairs (p_0) between X and Y varied from 1 to 5. Six tests were applied, including aSPC.dCor, aSPC.Sp, aSPC.P, permutation-based RV test, the Mantel test with the Euclidean distance and Pearson correlation, and dCov. One thousand datasets were simulated to calculate the empirical type I error and power. We used $B = 1000$ for any permutation-based tests. The simulation results are summarized in the left panel of Figure 4.2 with sample size $n = 40$.

First, the type I error rates were well controlled for all tests. Second, our aSPC.dCor test gave much higher power than the usual dCov test. For example, with only one truly associated pair, the power of aSPC.dCor was 86.5%, much higher than 12.0% of the dCov test. Third, due to the underlying non-monotonic true associations, as expected, none of the RV, aSPC.P, aSPC.Sp and Mantel tests performed well.

To further explore the performance of the tests with increasingly sparse associations, in addition to the above set-up with $p_0 = 5$, we added 75, 115, 195, 295 or 395 independent columns to matrix Y, each of which was simulated from a standard normal distribution. The power curves are shown in the right panel of Figure 4.2. It is clear that the power of aSPC.dCor remained significantly higher than that of the dCov test, whereas all other tests had no power.

4.5 Conclusions

We have proposed an adaptive and powerful association test called aSPC for two moderate- to high-dimensional random vectors. It has been shown to be more powerful in a a variety of simulations than several commonly used tests. In an application to a real genotype-gene expression dataset, under various moderately high dimensions for the SNPs and genes, the proposed test robustly and consistently gave more significant

P-values than other existing tests, which appeared to lose power dramatically for larger sets of the SNPs and genes. The proposed aSPC test can be regarded as a generalization of the standard RV test from low-dimensional data to moderate- to high-dimensional data with the incorporation of data-adaptive weighting on each variable pair. The main idea is that, for moderate- to high-dimensional data, often there will be many variable pairs that are not associated; treating these null pairs equally as other truly associated pairs will simply accumulate noises, leading to substantial power loss as in most other existing tests like the RV test. Hence, this main idea is related to the GEE-aSPU test in genetics. Indeed the aSPC test (more precisely, the version denoted aSPC.P with Pearson's correlation) is a special case of the GEE-aSPU test. However, due to its simplicity, the aSPC.P test has some computational advantage over the GEE-aSPU test, which in its currently implementation is not applicable to high-dimensional data. More importantly, the aSPC.P test can be easily extended by replacing the Pearson correlation coefficient with other coefficient, which may be more suitable for other non-linear associations. For example, if the distance correlation is used as in aSPC.dCor, it can detect non-monotonic associations. Compared to the usual dCov (or dCor) test, again due to its adaptiveness, the aSPC.dCor test is much more powerful for less dense or sparse signals for high-dimensional data, as shown in our simulations.

Various versions of the aSPC test are implemented in R package `aSPC`, freely available at `https://github.com/jasonzyx/aSPC`.

Table 4.1: The analysis results for the ADNI data. p and q denote the numbers of SNPs and of probes surviving the P-value cut-off based on the corresponding univariate SNP-gene expression associations.

Cut-off	(p, q)	SPC.P(γ)					aSPC.P	RV.asy	RV.perm	Mantel	dCov	aSPC.Sp	aSPC.dCor
		$\gamma = 1$	2	3	4	5-8, ∞							
1	(2483, 382)	5.85e-02	3.68e-02	2.87e-01	3.00e-04	1.00e-04	8.00e-04	6.89e-02	7.17e-02	8.69e-02	5.61e-02	9.00e-04	5.00e-04
0.9	(2274, 380)	3.63e-02	4.50e-02	2.16e-01	5.00e-04	1.00e-04	8.00e-04	6.47e-02	6.37e-02	9.76e-02	4.99e-02	8.00e-04	5.00e-04
0.8	(2069, 371)	2.29e-02	2.09e-02	2.22e-01	1.00e-04	1.00e-04	8.00e-04	3.87e-02	3.86e-02	7.40e-02	2.62e-02	8.00e-04	5.00e-04
0.7	(1871, 357)	3.26e-02	8.60e-03	2.39e-01	1.00e-04	1.00e-04	7.00e-04	2.01e-02	2.05e-02	5.81e-02	1.27e-02	8.00e-04	5.00e-04
0.6	(1647, 353)	1.39e-02	4.40e-03	1.74e-01	1.00e-04	1.00e-04	9.00e-04	9.22e-03	8.90e-03	4.71e-02	6.50e-03	7.00e-04	6.00e-04
0.5	(1435, 351)	1.62e-02	2.90e-03	2.80e-01	1.00e-04	1.00e-04	6.00e-04	6.69e-03	7.70e-03	5.96e-02	3.10e-03	6.00e-04	6.00e-04
0.4	(1228, 340)	1.99e-02	8.00e-04	2.54e-01	1.00e-04	1.00e-04	9.00e-04	1.91e-03	2.30e-03	1.49e-02	1.60e-03	9.00e-04	4.00e-04
0.3	(999, 306)	5.95e-02	1.20e-03	4.62e-01	1.00e-04	1.00e-04	8.00e-04	1.48e-03	1.50e-03	7.30e-03	9.00e-04	8.00e-04	1.00e-04
0.2	(756, 286)	7.54e-02	6.00e-04	5.93e-01	1.00e-04	1.00e-04	7.00e-04	6.07e-04	2.00e-04	2.00e-03	4.00e-04	9.00e-04	4.00e-04
0.1	(485, 245)	2.93e-01	1.00e-04	3.34e-01	1.00e-04	1.00e-04	7.00e-04	8.29e-05	3.00e-04	4.00e-04	2.00e-04	8.00e-04	4.00e-04

Table 4.2: Simulation I: empirical Type I error and power rates when the number of independent columns (denoted as "No. ind") is 25, 45, and 65 respectively. "RV.asy" and "RV.perm" stand for the asymptotic and permutation-based RV tests, respectively.

		SPC.P(γ)																
No.Ind		$\gamma = 1$	2	3	4	5	6	7	8	Inf	aSPC.P	aSPC.Sp	aSPC.dCor	RV.asy	RV.perm	Mantel	dCov	GEE-aSPU	MANOVA
25	Type I	0.047	0.053	0.049	0.050	0.050	0.051	0.052	0.053	0.054	0.046	0.049	0.049	0.053	0.055	0.050	0.052	0.055	0.046
	Power	0.417	0.844	0.886	0.932	0.917	0.908	0.879	0.852	0.589	0.933	0.893	0.828	0.840	0.838	0.098	0.819	0.955	0.378
45	Type I	0.055	0.052	0.052	0.052	0.052	0.053	0.050	0.050	0.048	0.052	0.049	0.050	0.052	0.051	0.050	0.053	0.061	0.045
	Power	0.196	0.538	0.587	0.753	0.732	0.759	0.710	0.700	0.425	0.749	0.674	0.587	0.539	0.538	0.074	0.522	0.832	0.174
65	Type I	0.056	0.055	0.050	0.052	0.054	0.050	0.050	0.051	0.049	0.050	0.050	0.047	0.056	0.055	0.052	0.054	0.057	0.041
	Power	0.118	0.352	0.371	0.581	0.558	0.627	0.576	0.578	0.328	0.594	0.506	0.450	0.355	0.354	0.072	0.345	0.702	0.110

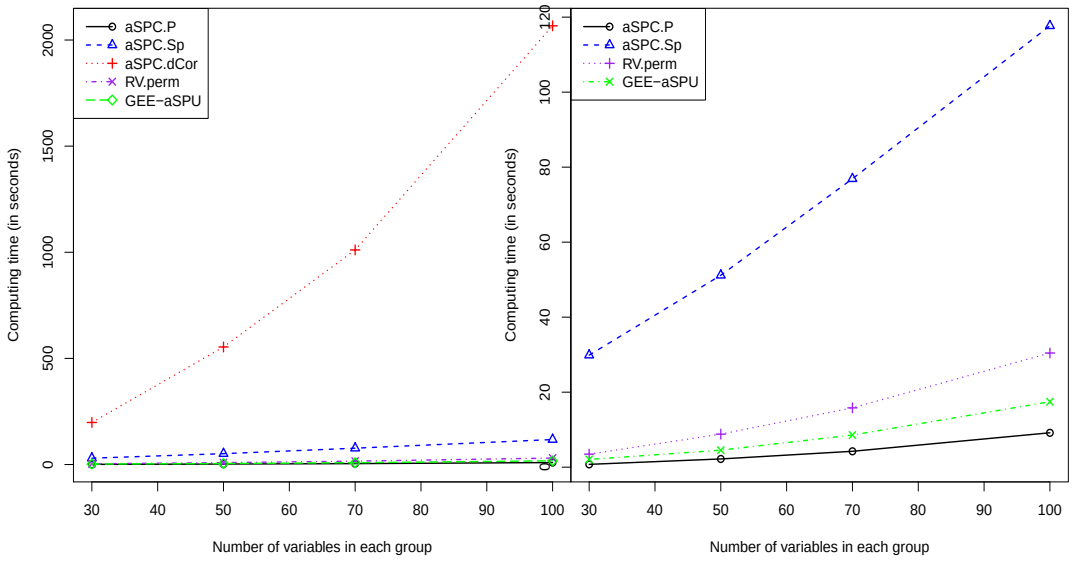


Figure 4.1: The computing time of the permutation-based RV, GEE-aSPU, aSPC.P, aSPC.Sp and aSPC.dCor tests. The left panel shows the computing time of aSPC.dCor test as compared to that of all the other tests, while the right panel is a zoom-in for all the tests except aSPC.dCor.

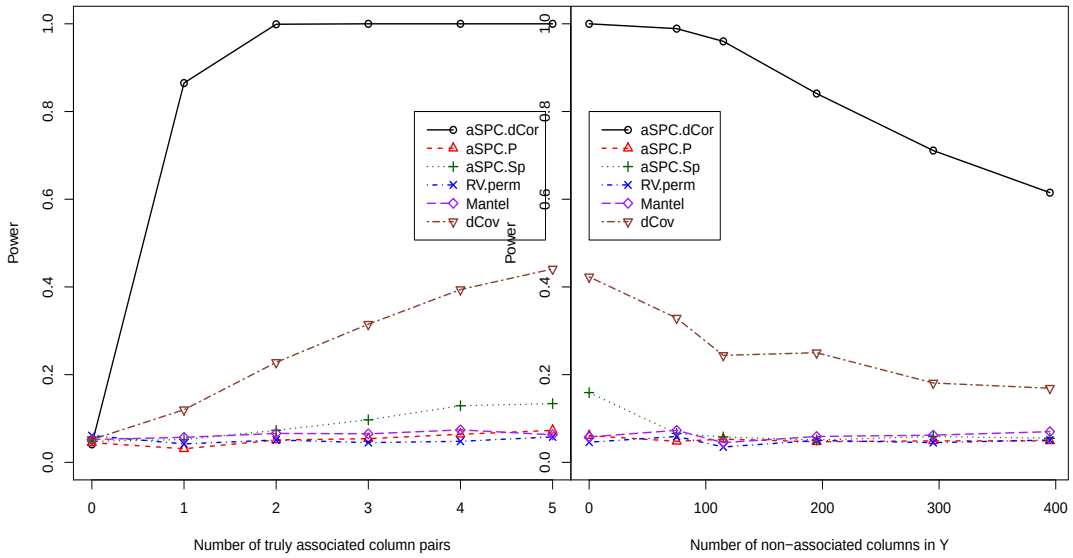


Figure 4.2: Simulation II results. The left panel: when the number of columns in X and Y are 5, the empirical type I error and power curves of the tests as the number of truly non-linearly associated column pairs between X and Y ranges from 0 (type I error) to 5. Right panel: when the number of non-linearly associated column pairs in X and Y is fixed at 5, the power curves of the tests as more and more non-associated columns are added to Y. The nominal significance level is 0.05.

Chapter 5

A powerful framework for integrating eQTL and GWAS summary data

5.1 Introduction

In spite of many successes, genome-wide association studies (GWAS) face two major challenges. The first is its limited statistical power even with tens to hundreds of thousands of individuals in a typical GWAS or mega-GWAS, thus missing many associated genetic variants, mostly single nucleotide polymorphisms (SNPs), due to the polygenic effects and small effect sizes. The second is that even for those few identified SNPs, since they often do not reside in protein-coding regions, it is difficult to interpret their function and thus biological mechanisms underlying complex traits. A new gene-based association test called *PrediXcan* was recently proposed to integrate GWAS individual-level data with an eQTL dataset, alleviating the above two problems in boosting statistical power of GWAS and facilitating biological interpretation of GWAS discoveries (Gamazon et al 2015). It was extended to GWAS summary association data (Torres et al 2017). A similar approach, called transcriptome-wide association study (TWAS), was proposed by another group for GWAS individual-level and summary data for one or more eQTL datasets (Gusev et al 2016). They are motivated by the key fact that many genetic

variants influence complex traits through transcriptional regulation. Focusing on the genetic component of expression avoids environmental noise influencing gene expression and complex traits, thus can increase statistical power. In addition, compared to standard GWAS, treating genes as analysis units reduces the number and thus burden of multiple tests, again leading to improved power. By applications to common diseases like T2D and complex traits like BMI, lipids and height, the authors have convincingly shown the power of integrating GWAS and eQTL data, gaining biological insights into complex traits. There are more follow-up studies applying TWAS to other diseases. For example, Gusev et al (2017) identified some new genes associated with schizophrenia; interestingly, they also confirmed a previous observation that, contrary to usual GWAS practice, the nearest gene to a GWAS hit often is not the most likely susceptibility gene, highlighting the critical role of incorporating gene expression to unravel disease mechanisms that may not be achieved by GWAS alone. The current standard and popular view is that PrediXcan and TWAS work because of their predicting or imputing *cis* genetic component of expression for a larger set of individuals in GWAS, facilitating the following expression-trait association testing. Based on this view, some new methods have been proposed to improve over TWAS by addressing some existing weaknesses in gene expression prediction (Bhutani et al 2017; Park et al 2017). In spite of its intuition and usefulness, the current view on PrediXcan and TWAS may not have told the whole story. Here we offer some new insights into PrediXcan and TWAS with a novel reformulation on their underlying association testing. Our key observation is that PrediXcan and TWAS share a common weighted association test; the weights on a set of SNPs in a gene are the *cis*-effects of the SNPs on gene expression (derived from an eQTL dataset). In other words, PrediXcan and TWAS put a higher weight on an SNP (eSNP) that is more strongly associated with the gene's expression level, in agreement with empirical evidence that eSNPs are more likely to be associated with complex traits and diseases (Nicholae et al 2010). This new formulation also points out the connection to existing weighted association analysis (Roeder et al 2006). More importantly, since the same association testing methodology in PrediXcan and TWAS suffers from power loss under some common situations, we develop an alternative and more powerful association test with broader applications. Since there is no uniformly most powerful gene-based association test, any single non-adaptive test will lose power in some situations; it is

important to develop and utilize adaptive tests to yield high power (Li and Tseng 2011; Lee et al 2012; Pan et al 2014). We propose using such an adaptive and powerful test under a general and rigorous framework of generalized linear models (GLMs), which can accommodate various types of quantitative, categorical and survival phenotypes and can adjust for covariates. It is applicable to both individual-level genotypic, phenotypic data and GWAS summary statistics. It si flexible to incorporate a single or multiple sets of weights derived from eQTL data or other data sources.

5.2 Methods

5.2.1 PrediXcan and TWAS

We briefly review PrediXcan and TWAS for GWAS individual-level data before giving our new formulation. One first builds a prediction model for a gene's expression level, called "genetically regulated expression (GReX)", by using the genotypes around the gene based on an eQTL dataset. Next, for a GWAS dataset, one uses the prediction model to predict or "impute" the GReX of the gene using the SNPs around the gene for each subject in the GWAS dataset. Specifically, for a given gene, suppose that in an eQTL dataset, Y^* and $X^* = (X_1^*, ..., X_p^*)'$ are the the expression level of and the p SNP genotype scores (with additive coding) around the gene. A linear model is assumed: $Y^* = \sum_{j=1}^p w_j X_j^* + \epsilon$, where w_j is the *cis*-effect of SNP j on gene expression and ϵ is the noise. Based on the eQTL dataset, one can use a method, e.g. elastic net (Zou and Hastie 2005) or a Bayesian linear mixed model (Zhou et al 2013) as used in PrediXcan and TWAS, to obtain the estimates $\hat{w}_j$'s. Now for a given GWAS dataset, for each subject i with the genotype scores $X_i = (X_{i,1}, ..., X_{i,p})'$ for the gene, the predicted GReX is $\widehat{\text{GReX}}_i = \sum_{j=1}^p \hat{w}_j X_{i,j}$. For a trait Y_i for subject i in the GWAS dataset, one simply applies a suitable GLM

$$g(E(Y_i)) = \beta_0 + \widehat{\text{GReX}}_i \beta_c = \beta_0 + \sum_{j=1}^{p} \hat{w}_j X_{i,j} \beta_c \tag{5.1}$$

to test for association between the trait and predicted/imputed expression with null hypothesis H_0: $\beta_c = 0$, where $g()$ is the canonical link function (e.g. the logit and the

identity functions for binary and quantitative traits respectively), and $E(Y_i)$ is the mean of the trait. One of the asymptotically equivalent Wald, Score and likelihood ratio tests can be used.

5.2.2 Novel reformulation and extensions

Here we first point out that PrediXcan and TWAS can be regarded as a special case of general association testing with multiple SNPs in a GLM:

$$g(E(Y_i)) = \beta_0 + \beta' X_i = \beta_0 + \sum_{j=1}^{p} X_{i,j}\beta_j. \tag{5.2}$$

The goal is to test $H_0 : \beta = (\beta_1, ..., \beta_p)' = 0$. It can be verified that both PrediXcan and TWAS are a weighted Sum test in the above general model (Pan 2009) with weights $\hat{w}_j$ on each SNP j; that is, PrediXcan and TWAS conduct the Sum test on H_0 with the genotype scores $X_{i,j}$ replaced by the weighted genotype scores $\hat{w}_j X_{i,j}$ in GLM (2). This new interpretation and formulation will facilitate our gaining insights into PrediXcan and TWAS, including their possible limitations, thus motivating some modifications for improvement. It offers a direct and intuitive justification for PrediXcan and TWAS: the two methods perform well due to their over-weighting on expression-associated SNPs (eSNPs), as supported by empirical evidence that eSNPs are more likely to be associated with complex traits and disease (Nicholae et al 2010). Obviously, it also suggests their extensions to other endophenotypes, and to incorporate prior knowledge and other data sources related to the GWAS trait of interest, such as previous linkage scans (Roeder et al 2006), though we do not pursue it here. More importantly, since the Sum test can be derived under the over-simplifying working assumption of $\beta_1 = \beta_2 = ... = \beta_p = \beta_c$ in (1) and (2) (i.e. all weighted SNPs have an equal effect size and the same effect direction, which is in general incorrect), we can see possible limitations of the Sum and thus PrediXcan and TWAS. As discussed in Pan (2009), Pan et al (2014) and others (Wu et al 2011), the Sum test may lose power if the effect directions of the (weighted) SNPs are different, or the effect sizes are sparse (i.e. with many 0s). Accordingly, one may apply other tests, e.g. the sum of squared score (SSU) test that is equivalent to a variance-component score test as used in kernel machine regression (also known as

SKAT in rare variant analysis) with a linear kernel and a nonparametric MANOVA (also called genomic distance-based regression) with the Euclidean distance metric (Wessel and Schork 2006), which may yield higher power under many situations (Pan 2011; Schaid 2010a, 2010b).

5.2.3 New method: aSPU

A class of the so-called sum of powered score (SPU) tests cover both the Sum and SSU tests as special cases. Specifically, we denote the unweighted and weighted score vectors for β in (1) as

$$U^* = (U_1^*, ..., U_p^*)' = \sum_{i=1}^{n} X_i'(Y_i - \hat{\mu}_i^0), \qquad U = (U_1, ..., U_p)' = WU^* = \sum_{i=1}^{n} WX_i'(Y_i - \hat{\mu}_i^0),$$

where $\hat{\mu}_i^0$ is the fitted mean of Y_i under H_0 (with $\beta = 0$) in (1), and $W = \text{Diag}(\hat{w}_1, ..., \hat{w}_p)$. The effects of the weights can be regarded as replacing the unweighted genotype scores $X_{i,j}$ by the weighted genotype scores $\hat{w}_j X_{i,j}$ in GLM (2). The Sum (i.e. PrediXcan and TWAS) and SSU tests based on the weighted genotypes are:

$$T_{\text{Sum}} = \sum_{j=1}^{p} U_j, \qquad T_{\text{SSU}} = U^T U = \sum_{j=1}^{p} U_j^2.$$

More generally, for an integer $\gamma \geq 1$, an SPU(γ) test is defined as

$$T_{\text{SPU}(\gamma)} = \sum_{j=1}^{p} U_j^{\gamma}.$$

It is clear SPU(1)=Sum and SPU(2)=SSU. Furthermore, for an even integer $\gamma \to \infty$, we have $T_{\text{SPU}(\gamma)} \propto \left(\sum_{j=1}^{p} |U_j|^{\gamma}\right)^{1/\gamma} \to \max_j |U_j| = T_{\text{SPU}(\infty)}$. The SPU($\infty$) is closely related to the minP test (but ignoring possibly varying variances of U_j's); often they performed similarly (Pan 2009).

Since there is no uniformly most powerful test, for a given situation, any non-adaptive test may or may not be powerful. By using various values of γ, we yield a class of SPU tests, one of which is expected to be more powerful in any given situation. For example, the Sum=SPU(1) test treats each SNP equally *a priori*, yielding

high power if all the SNPs are associated with the trait with similar effect sizes and the same association direction. On the other hand, when only a smaller subset of SNPs are associated with the trait, or their association directions are different, the SSU=SPU(2) test is often more powerful. As γ increases, SPU(γ) relies more on the SNPs that are more strongly associated with the trait, and is thus more powerful for more sparse association signals (i.e. fewer associated SNPs). In the end, as γ approaches ∞ (as an even integer), it only considers the most significant SNP.

Since the optimal value of γ is unknown and data-dependent, we propose using an **adaptive SPU (aSPU)** test to data-adaptively approximate the most powerful SPU test among a set of versatile SPU(γ) tests with various values of γ, thus maintaining high power in a wide range of scenarios. Empirically we have found that using $\Gamma = \{1, 2, 3, \ldots, 6, \infty\}$ often performs well and thus adopt it; the aSPU test is defined as

$$T_{aSPU} = \min_{\gamma \in \Gamma} P_{\mathrm{SPU}(\gamma)}, \tag{5.3}$$

where $P_{\mathrm{SPU}(\gamma)}$ is the p-value of the SPU(γ) test.

P-value calculations: Although asymptotic p-values for the SPU(1)=Sum and SPU(2)=SSU tests can be calculated (Pan 2009) (with possible small-sample adjustments (Lee et al 2012; Chen et al 2015; Wang 2016)), in general, we can use *one layer of Monte Carlo simulations* to estimate the p-values for all the SPU and aSPU tests *simultaneously* (Pan et al 2014). Specifically, we simulate null score vectors $U^{(b)} \sim N(0, V)$ for $b = 1, \ldots, B$, from its asymptotic null distribution as multivariate normal with mean 0 and covariance matrix V; there is a closed form solution for V (Pan et al 2014). Then the null statistics $T^{(b)}_{\mathrm{SPU}(\gamma)}$ can be calculated from the null score vectors $U^{*(b)}$ for $b = 1, \ldots, B$, and its p-value is $P_{\mathrm{SPU}(\gamma)} = [\sum_{b=1}^{B} I(|T^{(b)}_{\mathrm{SPU}(\gamma)}| \geq |T_{\mathrm{SPU}(\gamma)}|) + 1]/(B+1)$. Then the p-value for the aSPU test can be calculated as $P_{\mathrm{aSPU}} = [\sum_{b=1}^{B} I(T^{(b)}_{\mathrm{aSPU}} \leq T_{\mathrm{aSPU}}) + 1]/(B+1)$ with $T^{(b)}_{\mathrm{aSPU}} = \min_{\gamma \in \Gamma} p^{(b)}_{\gamma}$ and $p^{(b_1)}_{\gamma} = [\sum_{b \neq b_1} I(|T^{(b)}_{\mathrm{SPU}(\gamma)}| \geq |T^{(b_1)}_{\mathrm{SPU}(\gamma)}|) + 1]/B$.

5.2.4 Association testing with summary statistics

One practical way to increase the sample size is to form large consortia, aiming for meta analysis of multiple GWAS, for which often only summary statistics for single SNP-single trait associations, rather than individual-level genotypic and phenotypic

data, are available (and practically feasible for many cohorts with possibly different study designs). Hence it is extremely useful to develop methods like TWAS that are applicable to GWAS summary statistics as well as to GWAS individual-level data. The aSPU test is easily extended to GWAS summary statistics without individual-level data. Suppose that $Z_j = \hat{\beta}_j/\mathrm{SE}_j$ is the Z-statistic for association between the GWAS trait and SNP j, where $\hat{\beta}_j$ is the estimated (marginal and signed) association effect and SE_j is its standard error. We just need to simply redefine $U = WZ$ with $Z = (Z_1, Z_2, ..., Z_p)'$, then proceed as before. We use a reference sample (e.g. the 1000 Genome Project samples) to estimate linkage disequilibrium (LD) among the SNPs and thus the correlation matrix for Z and U (Kwak and Pan 2016; Gusev et al 2016).

5.2.5 Association testing with multiple sets of weights

Now we extend the aSPU test to the case with multiple sets of eQTL datasets, or more generally, multiple sets of weights. This is important because of the existence of multiple eQTL datasets measured from different populations or different tissues; it is in general unclear which one is most suitable. After applications with each eQTL dataset separately, it may gain statistical power and biological insights to combine the results across multiple eQTL datasets. Suppose we have K sets of weights, $W^{(k)} = \mathrm{Diag}(w_1^{(k)}, \ldots, w_p^{(k)})$ for $k = 1, 2, \ldots, K$, each estimated from a separate eQTL dataset or using a different method for a common eQTL dataset. To avoid the results depending on the varying scales of the sets of weights, we first standardize the weights to have $\sum_{j=1}^{p} |w_j^{(k)}| = 1$ for each k. Based on the score vector U^* (with individual-level data) or Z-statistics Z (with GWAS summary data) and the weights $W^{(k)}$, we define $U^{(k)} = W^{(k)}U^*$ or $U^{(k)} = W^{(k)}Z$ accordingly. As before, for a fixed γ, we first apply SPU(γ) to $U^{(k)} = (U_1^{(k)}, ..., U_p^{(k)})'$, yielding its test statistic $T_{SPU(\gamma;k)} = \sum_{j=1}^{p} (U_j^{(k)})^\gamma$ and and p-value $P_{SPU(\gamma;k)}$. We then Z-transform each p-value to a Z-statistic $z^*(\gamma; k) = \Phi^{-1}(1 - P_{SPU(\gamma;k)}/2)$, where $\Phi(.)$ is the CDF of a standard normal distribution. To recover the sign of each statistic, for an odd γ, we have $z(\gamma; k) = \mathrm{sign}(T_{SPU(\gamma;k)})z^*(\gamma; k)$; for an even γ or $\gamma = \infty$, we use $z(\gamma; k) = \mathrm{sign}(T_{SPU(1;k)})z^*(\gamma; k)$. We combine the K sets of weights through combining the K statistics $z(\gamma) = (z(\gamma; 1), ..., z(\gamma; K))'$ to form an

omnibus SPU(γ) test:

$$\text{SPU}(\gamma)\text{-O} = [z(\gamma) - \mu_0(\gamma)]' V^{-1}(\gamma)[z(\gamma) - \mu_0(\gamma)],$$

where $\mu_0(\gamma)$ and $V^{-1}(\gamma)$ are the mean vector and covariance matrix of $z(\gamma)$ under H_0, which can be calculated along with other p-values inside the single layer of simulations. Then, as usual, we combine the omnibus SPU(γ)-O tests into an omnibus aSPU test:

$$T_{\text{aSPU-O}} = \min_{\gamma \in \Gamma} P_{\text{SPU}(\gamma)\text{-O}},$$

where $P_{\text{SPU}(\gamma)\text{-O}}$ is the p-value of SPU(γ)-O. As before, the p-values of all the SPU(γ)-O and aSPU-O can be calculated in a single layer of Monte Carlo simulations.

It is easy to verify that SPU(1)-O is equivalent to the omnibus TWAS, denoted TWAS-O. Again, by combining SPU(1)-O and other SPU(γ)-O tests, we obtain the adaptive and omnibus aSPU-O test that may be more powerful across a wide range of scenarios.

5.3 Results

5.3.1 Application to the WTCCC data

We first applied the aSPU test and PrediXcan to the WTCCC individual-level data with the weights downloaded from the PrediXcan database, demonstrating the equivalence of the SPU(1) test and PrediXcan, and more importantly, that the aSPU test could identify more associated genes than PrediXcan in many cases. Specifically, first, following the same procedure of quality control (Burton et al 2007), we lifted the annotation of the WTCCC genotype data from hg18 to hg19 via the UCSC browser; second, we imputed the genotype data via the Michigan Imputation Server with the following specifications: 1000G Phase 1 v3 as the reference panel, SHAPEIT as the phasing algorithm and EUR (European) as the target population. After imputation, the variants with a minor allele frequency (MAF) > 0.05, the HWE exact test P-value > 0.05 and $\text{R}^2 > 0.8$ were kept. As Gamazon et al (2015), we kept only the HapMap Phase 2 subset of SNPs. We considered 7 traits/diseases, bipolar disorder (BD), coronary artery disease (CAD),

inflammatory bowel disease (CD), rheumatoid arthritis (RA), hypertension (HT), type 1 diabetes (T1D) and type 2 diabetes (T2D). The weights based on the DGN whole blood expression were downloaded from the PrediXcan database. There were 8917 genes whose expression levels could be predicted by elastic net with a cross-validated $R^2 > 0.01$; we thus tested on these 8917 genes with a conservative Bonferroni adjustment with a genome-wide significance level at $0.05/9000 = 5.56 \times 10^{-6}$.

As most of the genes are not expected to be significantly associated with a trait, we used a step-up procedure to increase the number of simulations when calculating the p-values of aSPU and aSPU-O in the subsequent data analysis. We started with a relatively small $B = 10^3$, and re-ran the tests with $B = 10^4$ for the genes with p-values $< 5 \times 10^{-3}$; we repeated this process by increasing B to 10 times of its previous value for the genes with p-values $< 5/B$) up to $B = 1 \times 10^7$; finally, to be more accurate for a p-value around the significance cut-off, we re-ran the tests on the genes with p-values between 10^{-5} and 10^{-6} with $B = 1 \times 10^8$.

Here are the main results. First, as shown in Figure 5.3, as expected, PrediXcan gave essentially the same results (i.e. p-values) as did the SPU(1) test for each of the seven traits. Hence, we treat the SPU(1) test to be equivalent to PrediXcan. Second, as shown in Figure 5.4, the aSPU test identified more significant genes than the SPU(1) test (or equivalently, PrediXcan) for traits CD, BD and T1D (i.e. (10, 3, 38) versus (8, 2, 29)), while it was the opposite for HT (i.e. 0 versus 1), and they were tied (with (1, 4, 0)) for CAD, RA and T2D; note the large difference for T1D. Table 5.1 lists the significant genes identified by the aSPU test but not by the SPU(1) test (and PrediXcan) at the genome-wide significance level; some of the significant genes were confirmed in later studies.

5.3.2 Application to the lipid GWAS summary data

We next applied our new methods and TWAS to a 2010 lipid GWAS summary dataset (∼100,000 samples, Teslovich et al 2010), while using its follow-up with a larger sample size (∼189,000) for partial validation. To facilitate comparison, we used the three sets of weights and the 1000 Genomes Project data as the reference sample, all downloaded from the TWAS database (on *Jan 11th, 2017*). The three sets of weights were based on three eQTL datasets: microarray gene expression data of peripheral blood from

1,245 unrelated subjects from the Netherlands Twin Registry (NTR) (Wright et al 2014), microarray expression data of blood from 1,264 subjects from the Young Finns Study (YFS), and RNA-seq measured in adipose tissue from 563 individuals from the Metabolic Syndrome in Men study (METSIM); for each pair of gene-eQTL dataset, we used the set of the optimal weights estimated by TWAS. For each trait, there were 1264, 3555 and 2295 significant cis-heritable genes with weights from the NTR, YFS and METSIM studies respectively, resulting in a total of 7114 genes being tested; when combining across three sets of weights, there were 1223 genes being tested. Thus, we used a conservative Bonferroni adjustment with $0.05/8500 = 5.88 \times 10^{-6}$ as the genome-wide significance level. The GWAS Z-scores were imputed for any missing SNPs using the IMPG algorithm (Pasaniuc et al 2014).

The new tests identified more associations

We numerically confirmed the equivalence between the SPU(1) test and TWAS (Figure 5.5). Hence, we used the results of the SPU(1) test to represent those of TWAS in the following. More importantly, the aSPU test could identify a larger number of significant genes than TWAS in every case across the four traits (HDL, LDL, TC and TG) and three sets of weights (NTR, YFS and METSIM); the same conclusion holds for the omnibus aSPU and omnibus TWAS tests (Table 5.2). As a partial validation, a high proportion of the identified genes covered at least one genome-wide significant SNP in the 2010 ($\sim$ 100,000 samples and the larger 2013 ($\sim$ 189,000 samples, Global Lipids Genetics Consortium 2013).

Compared to TWAS, the aSPU test can still maintain high power if many of the SNPs in a gene are not associated with a trait. For example, for trait HDL and gene DR1 with the YFS-based weights, among the 17 SNPs with non-zero weights, there were only one SNP with a p-value less than 5×10^{-7}, resulting in a non-significant p-value ($= 1.4 \times 10^{-4}$) by TWAS, or equivalently by SPU(1). Since an SPU(γ) test with a larger $\gamma > 1$ relied more on the SNPs with the smaller p-values (i.e. more strongly associated with the trait), it yielded a more significant p-value with a larger γ: the SPU(2) and SPU(5) tests gave p-value $= 2.9 \times 10^{-6}$ and 3.5×10^{-6} respectively, leading to the significant p-value $= 2.3 \times 10^{-6}$ of aSPU in the end. As shown in Figures 5.6-5.7, since the SPU(1) might not be the most powerful for a given gene, the aSPU test could

gain statistical power through more powerful other SPU tests like SPU(2).

The new tests identified new associations

Finally, we applied the aSPU and TWAS (and their omnibus versions) to the larger 2013 lipid dataset (Global Lipids Genetics Consortium 2013), listing the numbers of the significant genes identified by each method in Table 5.3. Again the aSPU test identified a much larger number of significant associations. The Manhattan plots for the pooled results of aSPU for each set of the weights and of aSPU-O combining the three sets of the weights for each trait are shown in Figures 1-2; a comparison between aSPU/aSPU-O versus TWAS/TWAS-O for trait LDL is shown in Figures 5.8-5.9. In total, aSPU and TWAS identified 17 and 14 new associations not overlapping with known risk loci respectively; among the 6 new associations uniquely identified by aSPU test, gene PFAS was reported to be associated with LDL in a later meta-analysis (Below et al 2016). The new associations identified by aSPU or/and TWAS are listed in Table 5.4 It is noteworthy that in Table 4, with the p-values close to the significance cut-off, the aSPU test barely missed the three significant genes uniquely identified by the SPU(1) test (i.e. TWAS); in contrast, the SPU(1) test gave the much larger p-values for several significant genes uniquely identified by aSPU.

It was shown previously that the aSPU test could control its type I error rate effectively in the context of unweighted association testing (Pan et al 2014), which is expected to hold in the current context. Nevertheless, we conducted a simulation study to confirm it. We used the individual-level (imputed) genotypic data of the WTCCC control and T2D samples with a combined sample size of $n = 4862$. We randomly generated a binary trait with an equal probability 0.5 for each subject, and calculated a summary Z statistic for each SNP. We then applied the aSPU test along with the asymptotic SPU(1) and SPU(2) tests to the individual-level data with the same PrediXcan-constructed weights based on the DGN whole blood gene expression data; in addition, we also applied the tests to the summary Z-statistics with the TWAS-constructed weights based on the NTR, YFS and METSIM gene expression data respectively; finally, we applied the omni-bus aSPU-O test to the summary statistics to combine results across the three sets of NTR, YFS and METSIM weights. As shown in the Q-Q plots in Figure 5.10, in each case each test controlled the type I error rate satisfactorily.

5.4 Conclusions

In summary, we have developed a powerful adaptive test (aSPU) to integrate GWAS and eQTL data. We have demonstrated its improved power over the existing methods; in fact, the same association test underlying the two existing methods, PrediXcan and TWAS, can be regarded as a special case of our proposed test, explaining why our proposed test may have improved power.

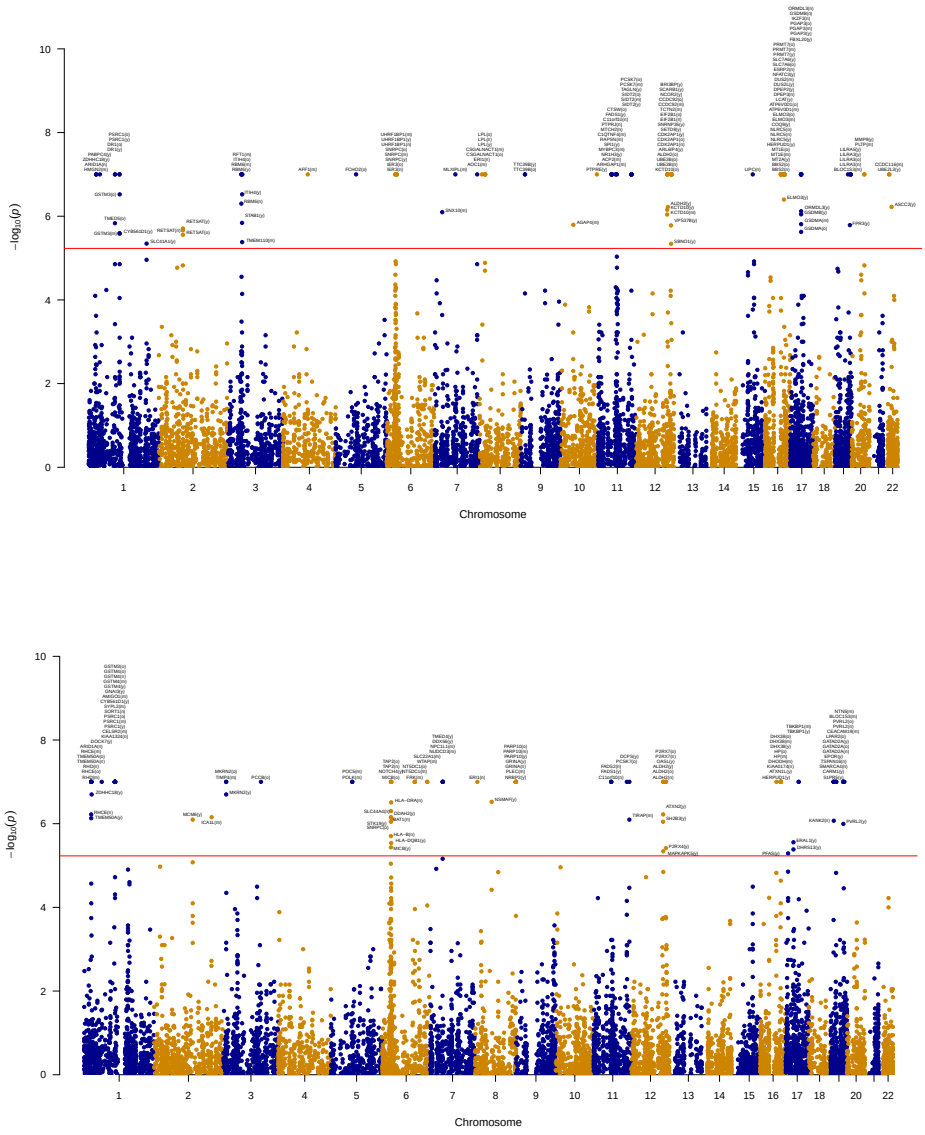


Figure 5.1: The Manhattan plots for the pooled results of aSPU and aSPU-O for traits HDL (top) and LDL (bottom) based on the 2013 lipid data. The letters "(n)", "(y)", "(m)" and "(o)" following a gene's name indicate the result of aSPU based on the NTR, YFS and METSIM weights and that of aSPU-O respectively.

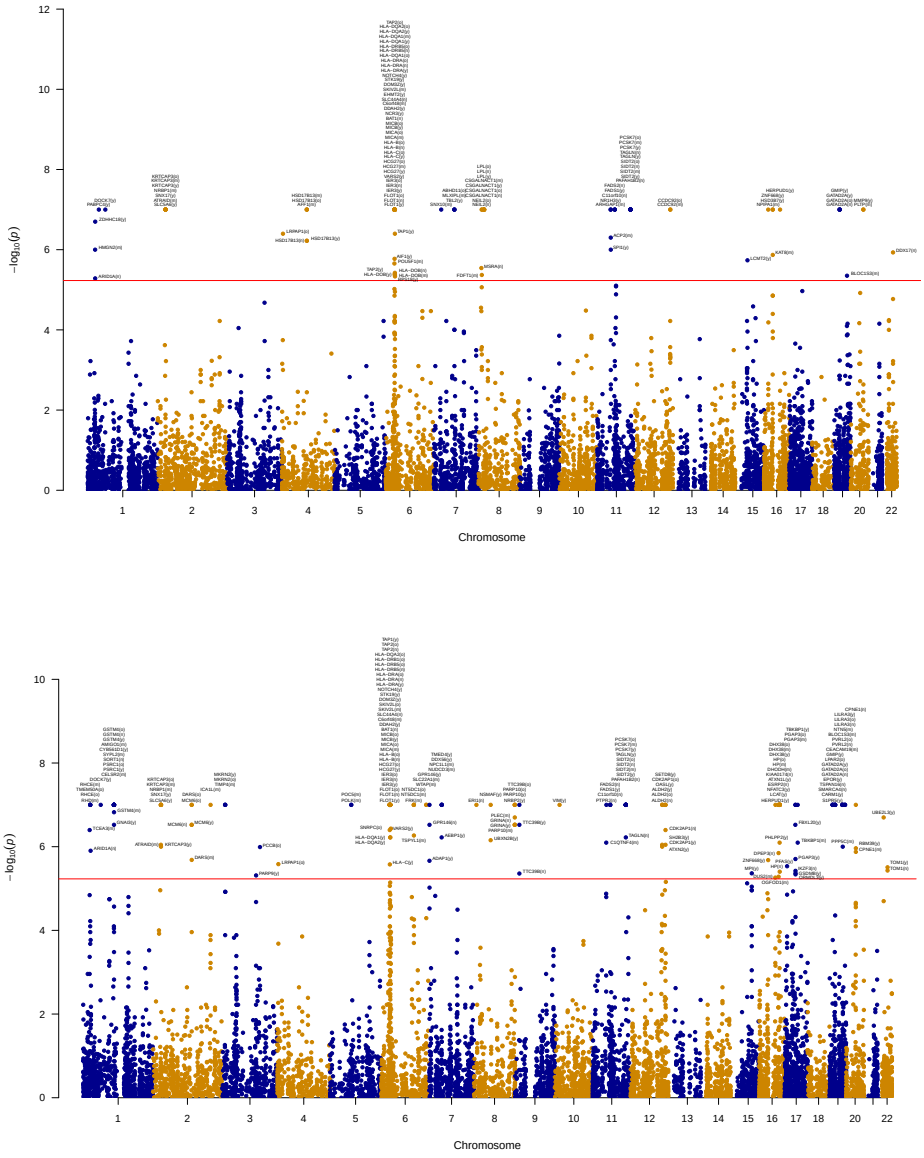


Figure 5.2: The Manhattan plots for the pooled results of aSPU and aSPU-O for traits TG (top) and TC (bottom) based on the 2013 lipid data.

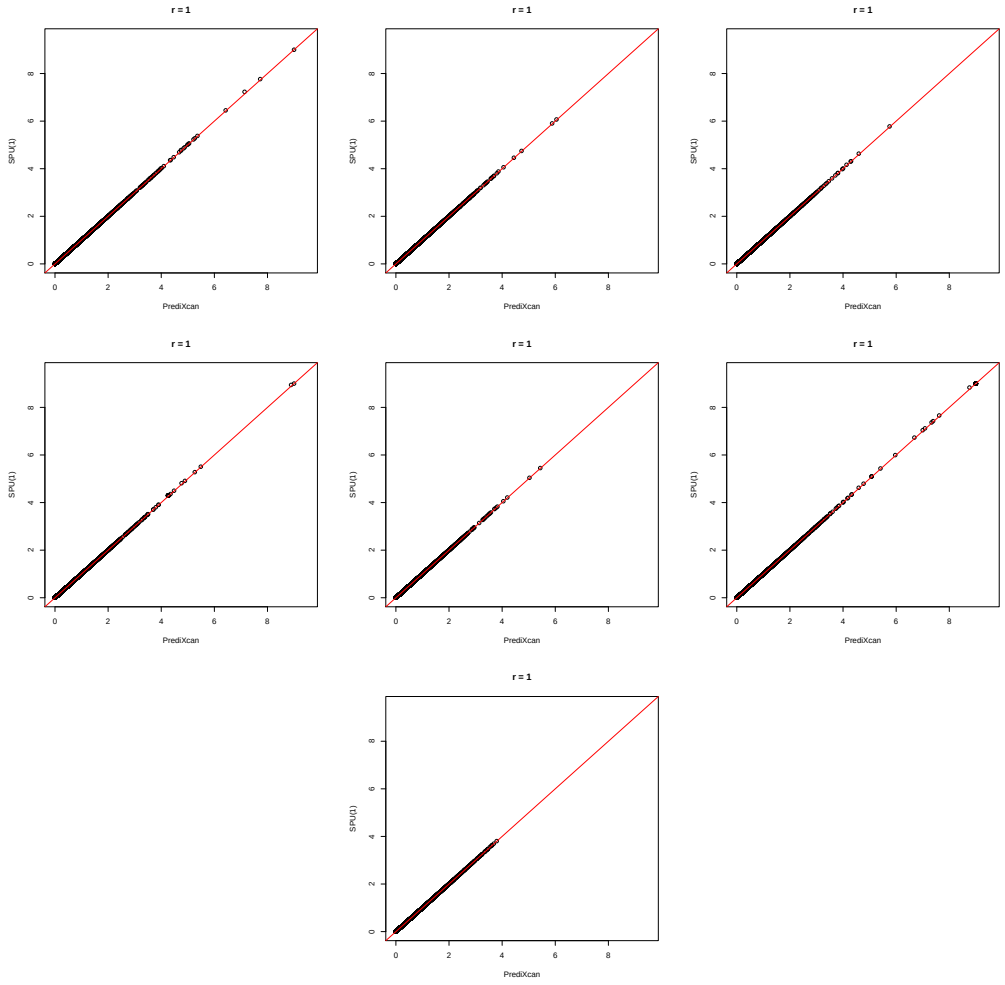


Figure 5.3: The scatter plots of the base 10 $-\log$(p-values) of the SPU(1) test and PrediXcan applied to the WTCCC data with weights derived from the DGN whole blood for traits CD, BD, CAD, RA, HT, T1D and T2D (from the left to the right and the top to the bottom). The Pearson correlation coefficient between the two sets of the -log p-values in each panel was equal to 1. For better visualization, all the p-values were truncated at 1×10^{-9}.

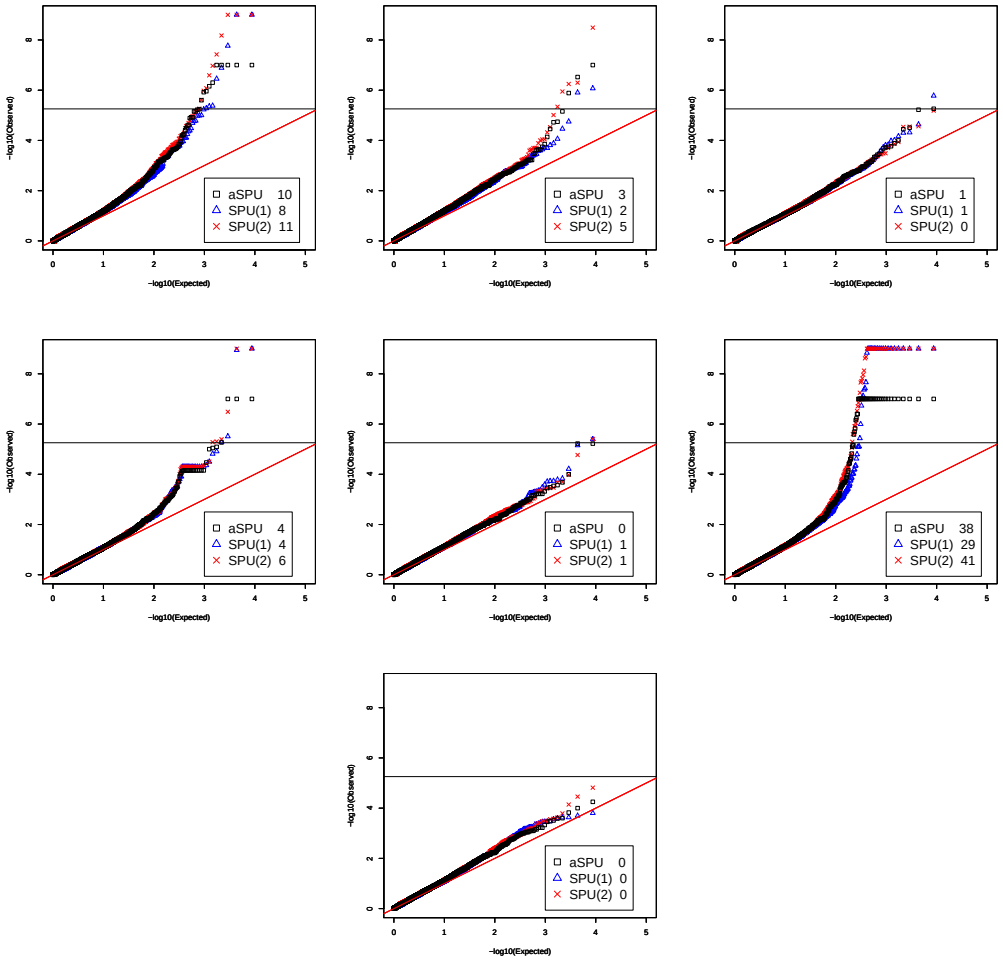


Figure 5.4: The Q-Q plots applied to the WTCCC data with the weights derived from the DGN whole blood gene expression for traits CD, BD, CAD, RA, HT, T1D and T2D (from the left to the right and the top to the bottom). The second column in each legend indicates the numbers of the significant genes identified at the genome-wide significance level of 5.56×10^{-6}. For better visualization, the p-values of aSPU were truncated at 1×10^{-7}, while those of the other two asympototic tests were truncated at 1×10^{-9}.

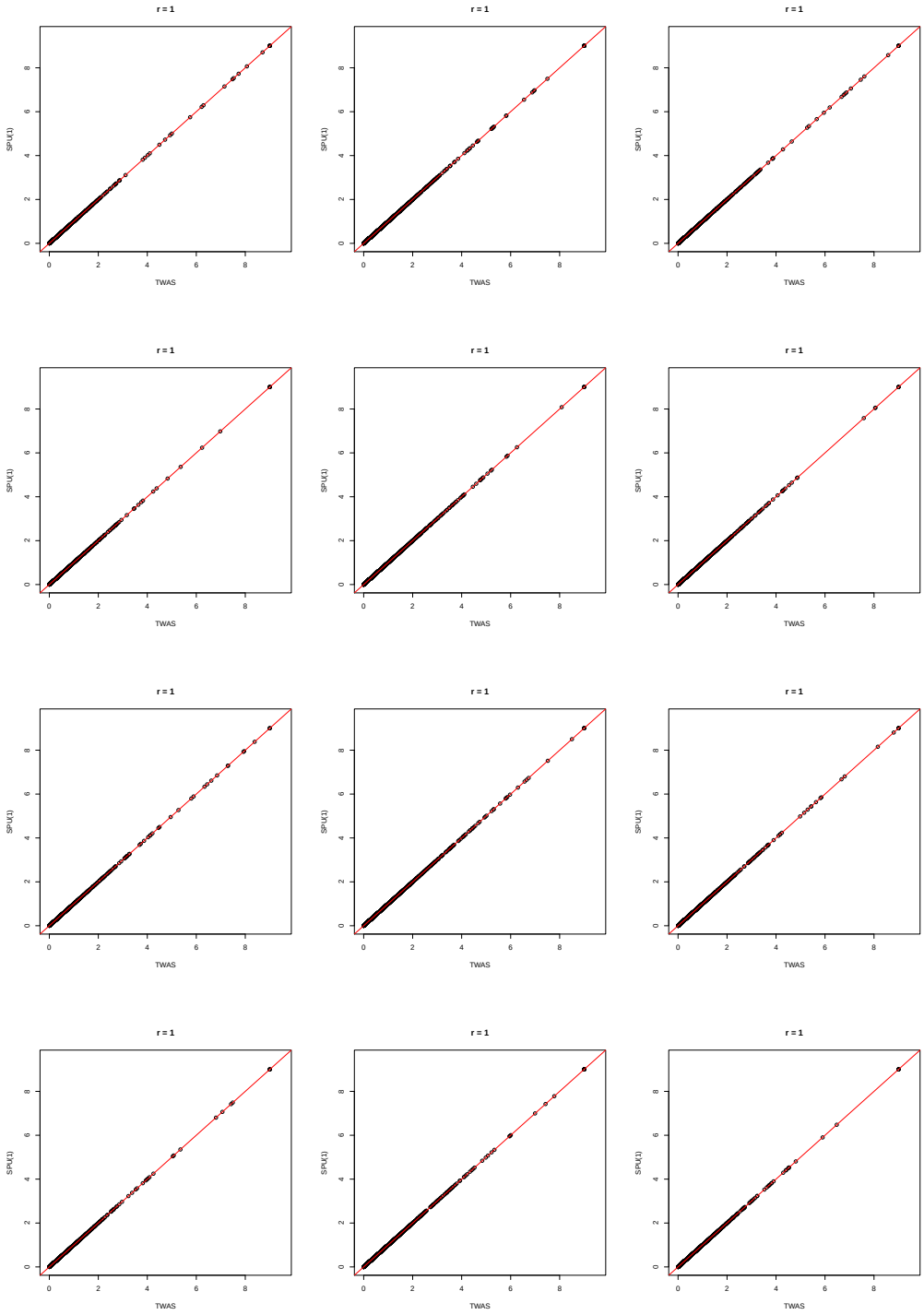


Figure 5.5: The scatter plots of the base 10 $-\log$(p-values) of the SPU(1) test and TWAS applied to 2010 lipid data with each set of the weights based on the NTR, YFS and METSIM studies, corresponding to the 1st, 2nd and 3rd columns, respectively. The panels from the top row to the bottom row correspond to traits HDL, LDL, TC and TG, respectively. The Pearson correlation coefficients in each panel was equal to 1. For better visualization, all the p-values were truncated at 1×10^{-9}.

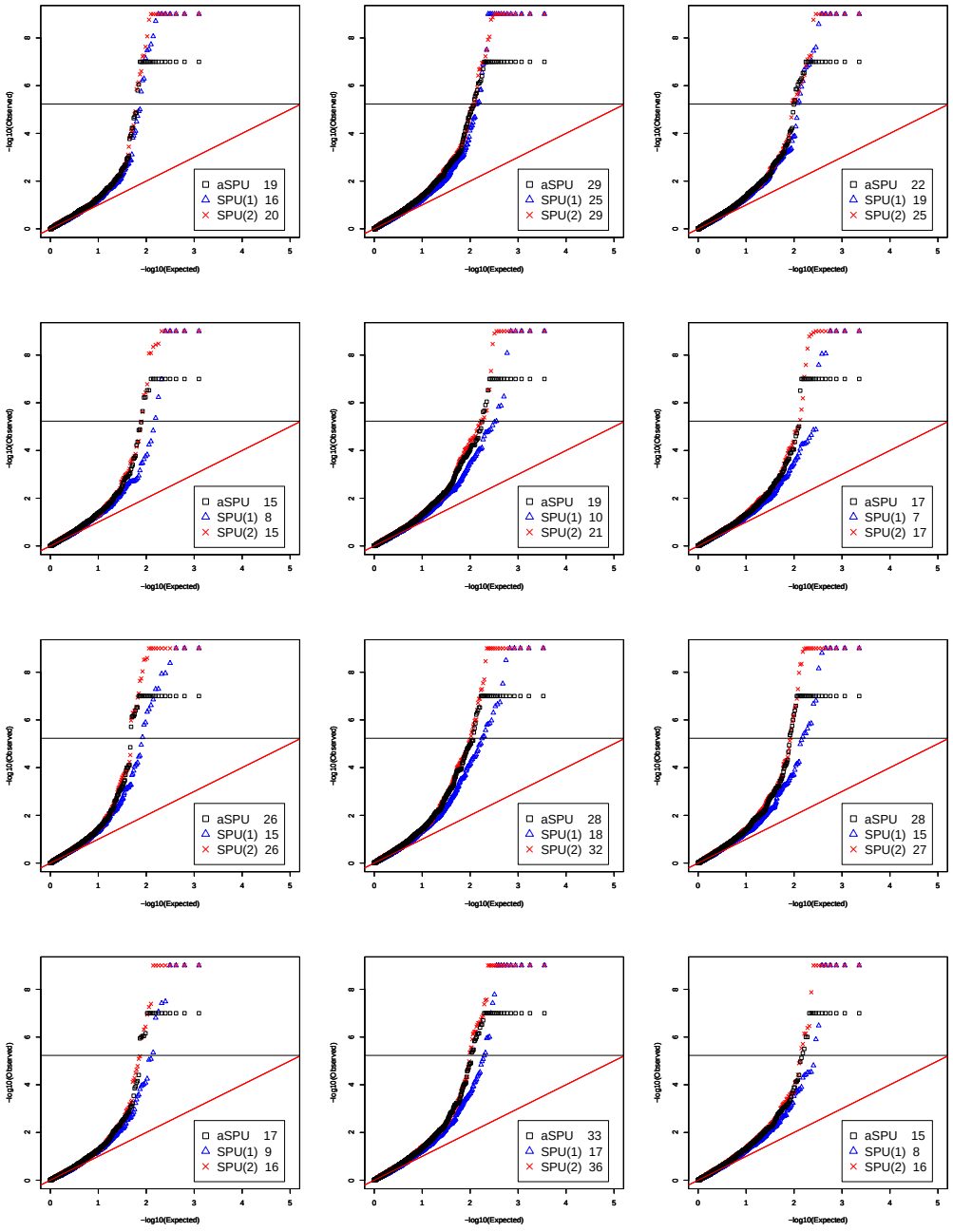


Figure 5.6: The Q-Q plots for the 2010 lipid data with each of the weights NTR, YFS and METSIM, corresponding to the 1st, 2nd and 3rd columns, respectively. The panels from the top row to the bottom row correspond to HDL, LDL, TC and TG, respectively. For better visualization, the p-values of aSPU were truncated at 1×10^{-7}, while those of the other two asympototic tests were truncated at 1×10^{-9}.

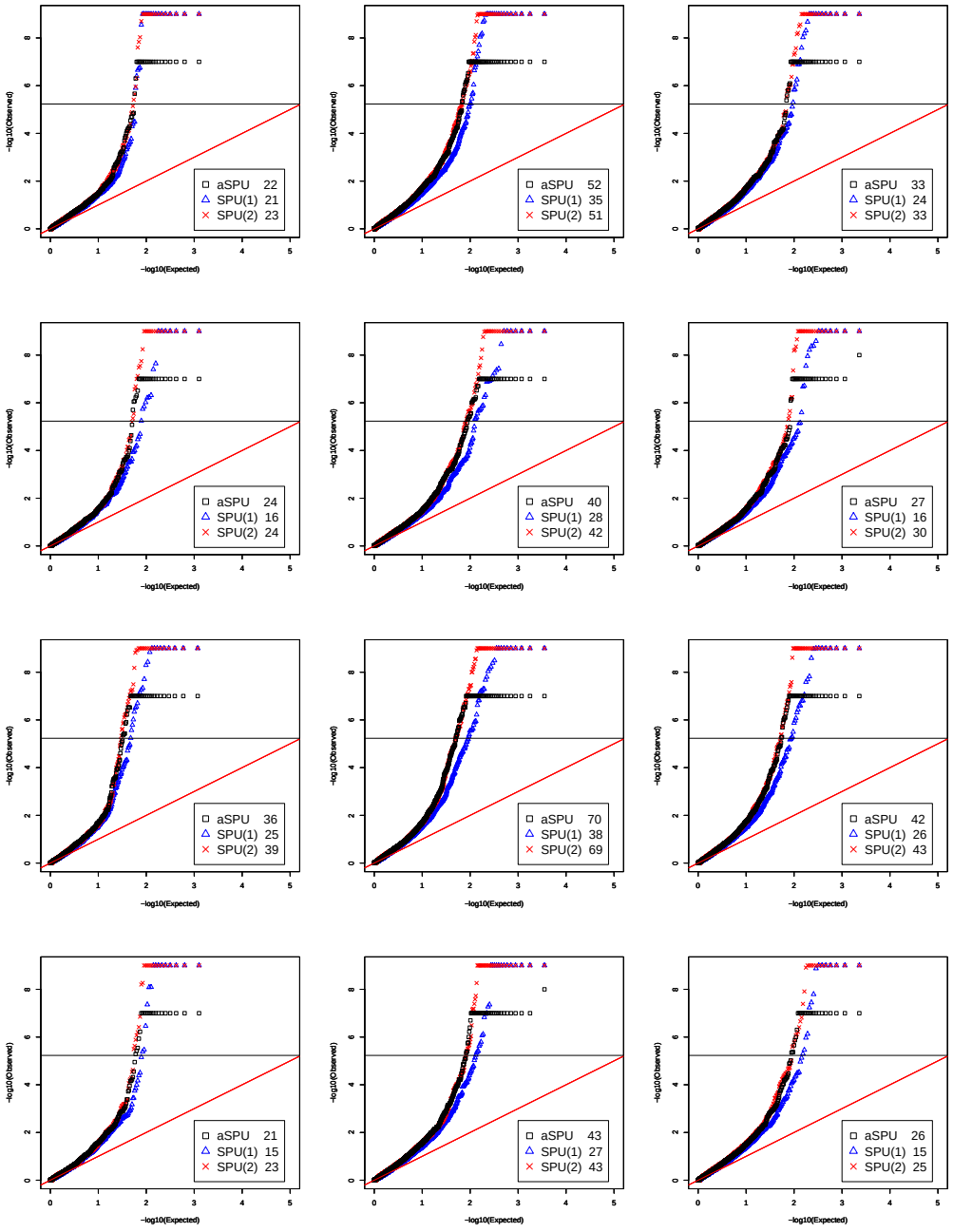


Figure 5.7: The Q-Q plots for the 2013 lipid data with each of the weights NTR, YFS and METSIM, corresponding to the 1st, 2nd and 3rd columns, respectively. The panels from the top row to the bottom row correspond to HDL, LDL, TC and TG, respectively. For better visualization, the p-values of aSPU were truncated at 1×10^{-7}, while those of the other two asympototic tests were truncated at 1×10^{-9}.

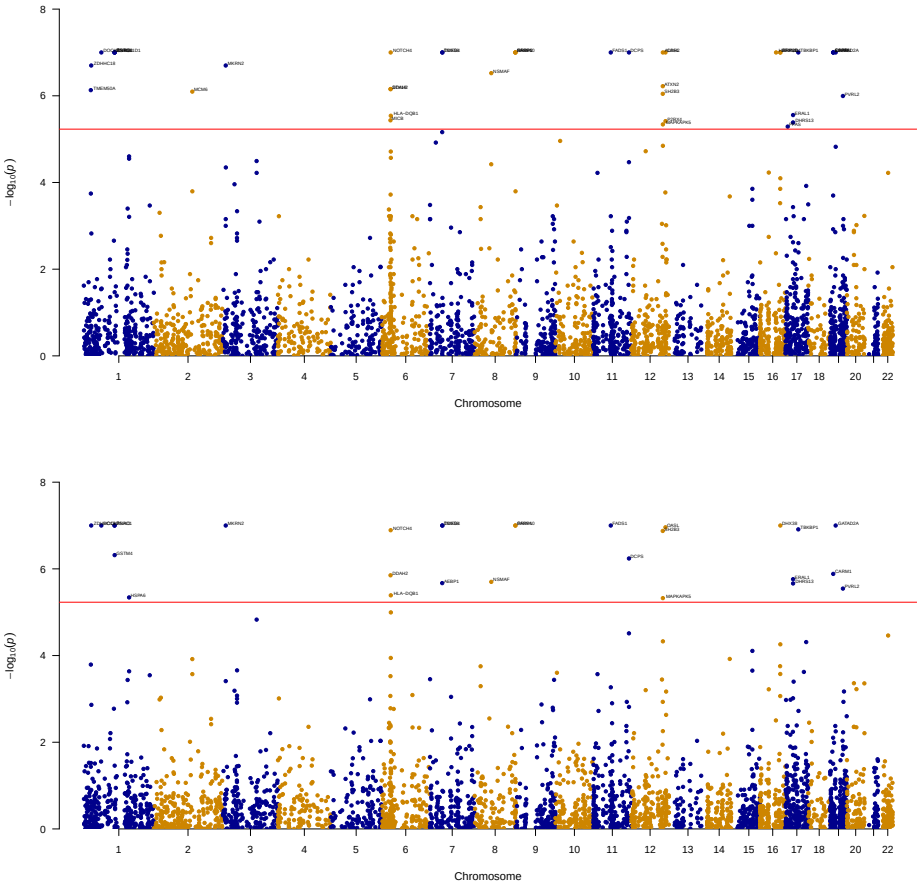


Figure 5.8: The Manhattan plots of aSPU (top) and TWAS (bottom) applied to the 2013 lipid dataset with trait LDL and the YFS-based weights. The p-values of both aSPU and TWAS were truncated at 1×10^{-7}.

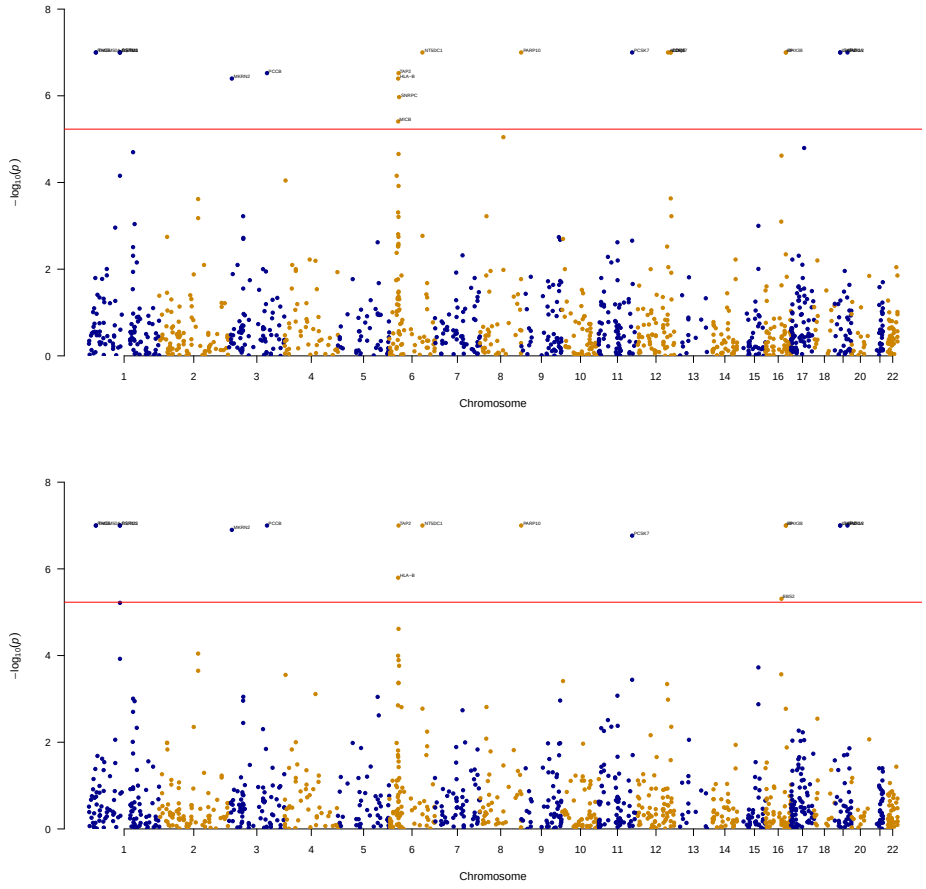
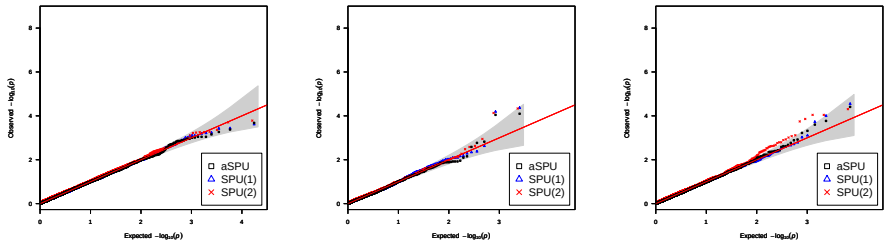
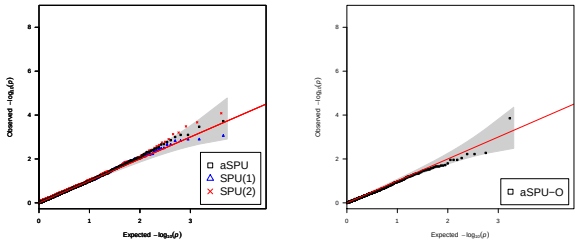


Figure 5.9: The Manhattan plots of aSPU-O (top) and TWAS-O (bottom) applied to the 2013 lipid dataset for trait LDL and combining over the three sets of the weights. The p-values of both aSPU-O and TWAS-O were truncated at 1×10^{-7}.



(a) (0.99, 0.98, 1.08) (b) (0.91, 0.98, 0.95) (c) (0.93, 0.93, 0.98)



(d) (0.94, 0.93, 0.92) (e) (0.84)

Figure 5.10: The Q-Q plots for the WTCCC genotypic data and a randomly simulated binary trait. (a) Results with the individual level data and the weights derived from the DGN whole blood gene expression; (b-d) the results with the summary statistics and the weights derived from the NTR, YFS and METSIM studies, respectively; (e) the results with the summary statistics and combining the three sets of the weights. For (a) - (d), the three numbers in each parenthesis correspond to inflation factors of aSPU, SPU(1) and SPU(2), respectively; for (e), the number in the parenthesis corresponds to the inflation factor of aSPU-O. For better visualization, the p-values of aSPU were truncated at 1×10^{-7}, while those of the other two asympototic tests were truncated at 1×10^{-9}.

Table 5.1: Significant genes identified by the aSPU test, but not by the SPU(1) test (or PrediXcan) at the genome-wide significance threshold of 5.56×10^{-6}. The validated gene-trait associations appeared in the following references: [1] Franke et al (2010); [2] Kenny et al (2012); [3] Plagnol et al (2011).

Trait	Gene	Chr.	SNPs in #SNPs	aSPU	SPU(1)	SPU(2)	Reported Valid. ref
CD	IRGM	5	34	5.0E-07	1.7E-04	3.8E-08	CD [1]
	P4HA2	5	7	2.5E-06	1.7E-01	8.4E-07	CD [2]
	PTGER4	5	22	1.1E-06	1.0E-05	2.6E-07	CD [1]
	RBM22	5	15	7.0E-07	1.4E-02	5.5E-06	-
BD	JAKMIP1	4	21	1.0E-07	1.8E-05	3.2E-09	-
CAD	PDK1	2	33	5.5E-06	5.8E-03	2.7E-05	-
T1D	ALDH2	12	19	1.0E-07	6.6E-05	1.1E-07	-
	BCL2L15	1	21	1.0E-07	5.1E-04	2.1E-06	T1D [3]
	HFE	6	61	7.0E-07	5.8E-01	5.7E-08	-
	MPHOSPH10	2	20	6.0E-07	1.2E-01	1.4E-07	-
	PGBD1	6	33	1.0E-07	3.0E-04	2.4E-09	-
	PRSS16	6	32	1.5E-06	6.3E-05	2.0E-07	-
	TMEM116	12	23	2.0E-06	5.0E-04	6.9E-07	-
	ZNF193	6	64	1.0E-07	2.3E-02	5.0E-11	-
	ZSCAN12	6	14	4.0E-07	8.1E-06	2.5E-06	-

Table 5.2: The numbers of the significant genes identified by analyzing the 2010 lipid data for each single set of the weights and the combined one (i.e. with the omnibus aSPU and TWAS tests). The numbers a/b/c in each cell indicate the numbers of (a) the significant genes; (b) the significant genes that covered a genome-wide significant SNPs in the 2010 lipid data; (c) the significant genes that covered a genome-wide significant SNPs in the 2013 lipid data.

Trait	Test	NTR	YFS	METSIM	Combined
HDL	aSPU	19/16/17	29/27/29	22/19/22	21/17/17
	TWAS	16/14/15	25/22/24	19/15/19	20/16/17
LDL	aSPU	15/15/15	19/18/18	17/16/17	14/13/13
	TWAS	8/7/8	10/9/9	7/7/7	7/7/7
TG	aSPU	17/16/17	33/30/32	15/14/14	20/19/19
	TWAS	9/9/9	17/16/17	8/7/7	12/11/11
TC	aSPU	26/25/26	28/26/27	28/28/28	20/20/20
	TWAS	15/14/15	18/16/17	15/14/15	14/13/13

Table 5.3: The numbers of the significant genes identified by analyzing the 2013 lipid data for each single set of the weights and the combined one (i.e. with the omnibus aSPU and TWAS tests). The numbers a/b/c in each cell indicate the numbers of (a) the significant genes; (b) the significant genes that covered a genome-wide significant SNPs in the 2010 lipid data; (c) the significant genes that covered a genome-wide significant SNPs in the 2013 lipid data.

Trait	Test	NTR	YFS	METSIM	Combined
HDL	aSPU	21/18/21	52/39/48	33/24/32	31/21/29
	TWAS	21/17/20	35/26/33	24/17/23	26/16/24
LDL	aSPU	24/23/24	40/34/37	27/24/26	20/17/18
	TWAS	16/14/16	28/23/25	16/14/15	17/15/16
TG	aSPU	21/19/21	43/37/42	26/20/25	23/21/23
	TWAS	15/13/15	27/22/25	15/12/15	16/15/16
TC	aSPU	36/27/35	70/52/66	42/36/42	37/32/36
	TWAS	25/18/24	38/28/35	26/21/26	25/20/23

Table 5.4: Significant gene-trait associations identified by aSPU or/and TWAS with no known risk loci within 500kb. The column "Sig.test" indicates the corresponding association was detected by aSPU or SPU(1) or both.

Trait	Weight	Gene	Chr.	Locus start	Locus end	aSPU	SPU(1)	SPU(2)	Sig.test
HDL	YFS	SLC41A1	1	205758221	205782876	4.5E-06	7.9E-04	2.2E-04	aSPU
HDL	YFS	ASCC2	22	30184597	30234271	6.0E-07	9.7E-04	6.2E-06	aSPU
LDL	YFS	PFAS	17	8150936	8173809	5.1E-06	4.2E-03	1.1E-06	aSPU
TG	METSIM	ARHGAP1	11	46698630	46722165	1.0E-07	3.8E-04	1.2E-08	aSPU
TC	YFS	ZNF668	16	31072164	31085641	2.1E-06	3.7E-05	1.4E-05	aSPU
TC	YFS	PFAS	17	8150936	8173809	2.9E-06	5.4E-03	1.8E-06	aSPU
LDL	YFS	HSPA6	1	161494036	161496681	2.8E-05	4.6E-06	3.2E-02	SPU(1)
TG	YFS	PACS1	11	65837834	66012218	8.2E-06	5.7E-06	1.2E-06	SPU(1)
TC	YFS	ADCY3	2	25042038	25142708	1.1E-05	4.1E-06	2.5E-02	SPU(1)
HDL	NTR	RETSAT	2	85569078	85581848	2.2E-06	1.3E-06	2.3E-06	Both
HDL	YFS	RETSAT	2	85569211	85581743	2.0E-06	2.3E-06	1.5E-06	Both
HDL	YFS	PTPRE	10	129705325	129884119	1.0E-07	3.7E-08	1.4E-01	Both
HDL	METSIM	SNX10	7	26331541	26413949	8.0E-07	5.6E-07	8.7E-07	Both
LDL	YFS	ERAL1	17	27181956	27188085	2.8E-06	1.7E-06	2.2E-06	Both
LDL	YFS	DHRS13	17	27224799	27230089	4.1E-06	2.2E-06	2.4E-06	Both
LDL	METSIM	ICA1L	2	203640690	203736708	7.0E-07	2.1E-07	5.9E-07	Both
TG	YFS	LCMT2	15	43619974	43622803	1.8E-06	1.6E-06	3.2E-06	Both
TC	NTR	ARID1A	1	27022521	27109023	1.2E-06	9.7E-07	1.4E-06	Both
TC	YFS	PARP9	3	122246771	122283424	4.9E-06	2.4E-06	3.7E-06	Both
TC	YFS	MPI	15	75182346	75191798	4.3E-06	6.1E-07	6.8E-07	Both

Chapter 6

Discussion and future work

As it is not feasible to apply the existing score-based testing under complex models due to either the lack of a close-form solution of score vectors or no software output, Chapter 2 introduced a method to approximate the score vector by MLEs. Our proposed general approach is a two-step procedure. In the first step, a full model including a set of parameters to be tested is fitted, then in the second step the score vector for the parameter and its covariance matrix are approximated based on the parameter estimates and their covariance matrix. In this way, a score-based test can be applied without directly calculating the score vector (and its covariance matrix), which may not be easy to derive based on existing software packages, such as for mvLMM and GLMM. Due to the nature of the proposed two-step approach, the validity of the approach depends on both the first step and the asymptotics. For example, if we have a familial dataset with trait-ascertained samples, then it is necessary to appropriately account for the sample ascertainment in step one, e.g., based on some family-based association testing procedures [Moerkerke et al., 2010; Zhang et al., 2012]. Furthermore, because the proposed approximation to the score vector is based on the asymptotics of the parameters to be tested, it has some limitations. First, if the sample size is too small or more generally, if the conditions for the asymptotics do not hold, e.g., in analysis of rare variants [Chen et al., 2013; Jiang et al., 2014], then it may not perform well with inflated false positives and false negatives. Second, in order to obtain a point estimate of the parameters to be tested, a full model has to be fitted, which may not be even computationally feasible if the number of the parameters to be estimated is too large relative to the

sample size. Nevertheless, the proposed method offers a simple and practical way to extend many score-based tests to more complex models, for which the score vector is either unavailable from software or has no closed-form solution. In addition to testing for main effects as considered here, it will also be interesting to explore the use of the proposed method to detect gene-gene and gene-environment interactions [Tzeng et al., 2011]. For example, the Minnesota Twin Family Study (MTFS) is a longitudinal study of twins and their parents to examine factors to the etiology of substance abuse and related problelms [Lacono and McGue, 2002]. An application of our proposed method would be to study the associations between clinical phenotype (e.g nicotine factor, alcohol consumption factor, drug factor etc.) and genetic variations accounting for both genetic and (shared) environmental factors. There are several challenging while very interesting problems of modeling this family data. First, as twins families share not only genetic but also environment, both genetic and shared environmental factors need to be taken into consideration. Second, as MTFS is a longitudinal study, each phenotype (e.g. nicotine factor) has multiple time points. Thus, how to model longitudinal phenotype while accounting for genetic and shared environment factors is an interesting question. Third, as some of phenotypes might be correlated, it may help to gain power to consider multiple traits in association analysis. Li et al. [2011] proposed a generalized least squares approach for genome-wide quantitative trait association analysis in families, which is a feasible and efficient model providing both genetic and environmental variance estimates. Instead of modeling the clinical phenotype at a single time point, we can treat the longitudinal phenotype as multiple phenotypes and fit a multivariate linear mixed model (e.g. by `GEMMA`) by supplying the genetic and environmental variance estimated from RFGLS. Another way to model the twins family data is that we can model multiple phenotypes (e.g. nicotine factor, drug factor and alcohol consumption factor) by linear mixed-effects model approaches. At the end, based on the MLE by fitting the mixed-effects models, we can construct the approximated score vectors accordingly and conduct score-based testing (e.g aSPU test).

In Chapter 3, we showed that potential power gain could be achieved by accounting genetic heterogeneity. As integrative analysis of multiple types of data may provide more information to capture the disease heterogeneity, our future work could develop a novel machine learning method to discover the subpopulations for a disease of interest. For

example, ADNI data consists of structural and functional magnetic resonance imaging (sMRI and fMRI), single-nucleotide polymorphism (SNP)/genotyping and messenger ribonucleic acid (mRNA) etc. Rather than considering unknown genetic genterogeneity here, we may utilize mRNA data, for example, to identify the subpopulations in patient group and conduct association testing accounting on the identified "known" subpopulation, which may increase power.

Motivated by a reformulation of RV test, we proposed a new method aSPC to test association between two random vectors in Chapter 4. In the current implementation of the new tests, we have resorted to permutations to calculate their P-values, which seems feasible and satisfactory in many applications. However, it would be interesting to establish their asymptotics as both p and q diverge with n (Xu et al 2016), which may be challenging due to the dependencies among the individual correlation coefficients in each SPC test statistic. Nevertheless, an asymptotic theory will be useful in facilitating speedy P-value calculations, especially for a high significance level.

Chapter 5 introduced a statistical method integrating eQTL and GWAS (summary or individual level) data. Importantly, our new formulation of PrediXcan and TWAS suggests other possible extensions, e.g. applications not only to other informative endophenotypes, but also to incorporate other sources of information like previous linkage scans (Roeder et al 2006) and multiple phenotypes (Kim et al 2015; Zhu et al 2015), which will be investigated in the future.

Chapter 7

References

1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. Nature 467: 1061-1073.

Bates D, Maechler M, Bolker B, Walker S. 2014. lme4: linear mixed-effects models using Eigen and S4. R package version 1.1-6. `http://CRAN.R-project.org/package=lme4`

Balding DJ. 2006. A tutorial on statistical methods for population association studies Nat Rev Genet 7: 781-791.

Boos DD. 1992. On generalized score test. Am Stat 46: 327-333.

Below JE, Parra EJ, Gamazon ER et al. 2016. Meta-analysis of lipid-traits in Hispanics identifies novel loci, population-specific effects, and tissue-specific enrichment of eQTLs. Scientific Reports 6.

Breslow NE, Clayton DG. 1993. Approximate inference in generalized linear mixed models. J Am Stat Assoc 88: 9-25.

Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, Kwiatkowski PD, McCarthy IM, Ouwehand HW, Samani JN, and others. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447(7145): 661-678.

Bhutani K, Sarkar A, Park Y, Kellis M, Schork NJ. 2017. Modeling prediction error improves power of transcriptome-wide association studies. bioRxiv. doi: http://dx.doi.org/10.1101/108316.

Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, Kwiatkowski PD, McCarthy IM, Ouwehand HW, Samani JN, Todd AJ, Donnelly P. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661-678.

Colantuoni C, Lipska BK, Ye T, et al. 2011. Temporal dynamics and genetic control of transcription in the human prefrontal cortex. Nature 478(7370): 519-523.

Chen J, Chen W, Zhao N, Wu MC, Schaid DJ. 2015 Small Sample Kernel Association Tests for Human Genetic and Microbiome Association Studies. Genet Epidemiol 40: 5-19.

Chen H, Meigs JB, Dupuis J. 2013. Sequence kernel association test for quantitative traits in family samples. Genet Epidemiol 37: 196-204.

Cho YS, Chen CH, Hu C, Long J, Ong RTH, Sim X, Takeuchi F, Wu Y, Go MJ, Yamauchi T and others. 2012. Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. Nat Genet 44(1): 67-72.

Clayton D, Chapman J, Cooper J. 2004. Use of unphased multilocus genotype data in indirect association studies. Genet Epidemiol 27(4): 415-428.

Darabi H, Humphreys K. 2011. Single- and multi-locus association tests incorporating phenotype heterogeneity. Hum Hered 71(1): 11-22.

Diggle P, Heagerty P, Liang KY, Zeger S. 2013. Analysis of Longitudinal Data. Oxford University Press.

Escoufier Y. 1970. Echantillonnage dans une population de variables aléatoires réelles. Department de math.; Univ. des sciences et techniques du Languedoc.

Escoufier Y. 1973. Le traitement des variables vectorielles. Biometrics 29: 751-760.

Fan R, Knapp M. 2003. Genome association studies of complex diseases by case-control designs. Am J Hum Genet 72: 850-868.

Fan R, Chiu C, Jung J, Weeks DE, Wilson AF, Bailey-Wilson, JE, Amos CI, Chen Z, Mills JL, Xiong M. 2016. A Comparison Study of Fixed and Mixed Effect Models for Gene Level Association Studies of Complex Traits. Genet Epidemiol 40: 702-721.

Franke A, McGovern DP, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, Lees CW, Mathew CG, Montgomery GW, Prescott NJ and others. 2010. Genome-wide meta- analysis increases to 71 the number of confirmed Crohns disease susceptibility loci. Nat Genet 42(12):1118-1125.

Fuchsberger C, Abecasis GR and Hinds DA. 2015 minimac2: faster genotype imputation. Bioinformatics 31: 782-784.

Gamazon ER, et al. 2015 A gene-based association method for mapping traits using reference transcriptome data. Nat Genet 47: 1091-1098.

Global Lipids Genetics Consortium. 2013 Discovery and refinement of loci associated with lipid levels. Nat Genet 45: 1274-1283.

GTEx Consortium. 2015. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 348: 648-660.

Gusev, A. et al. 2016. Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet 48: 245-252.

Gusev A, Mancuso N, Finucane HK, Reshef Y. et al. 2017. Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. bioRxiv. doi: https://doi.org/10.1101/067355.

Hartig M, Truran-Sacrey D, Raptentsetsang S, Simonson A, Mezher A, Schuff N, Weiner M. 2012. UCSF FreeSurfer Overview and QC Ratings. San Francisco: Alzheimers Disease Neuroimaging Initiative (ADNI).

Hara K, Fujita H, Johnson TA, Yamauchi T, Yasuda K, Horikoshi M, Peng C, Hu C, Ma RC, Imamura M and others. 2014. Genome-wide association study identifies three novel loci for type 2 diabetes. Hum Mol Genet 23(1): 239-246.

Howie B, Fuchsberger C, Stephens M, Marchini J and Abecasis GR. 2012. Fast and accurate genotype imputation in genome-wide association studies through prephasing. Nat. Genet 44: 955-959.

Hua WY, Ghosh D. 2015. Equivalence of Kernel Machine Regression and Kernel Distance Covariance for Multidimensional Phenotype Association Studies. Biometrics 71: 812-820.

He Z, Zhang M, Lee S, Smith JA, Guo X, Palmas W, Kardia SLR , Diez Roux AV, Mukherjee B. (2015). Set-based tests for genetic association in longitudinal studies. Biometrics 71: 606-615.

Iacono WG, McGue M. Minnesota twin family study. 2002. Twin Research 5(05): 482-487.

Li X, Basu S, Miller MB, Iacono WG, McGue M. 2011. A rapid generalized least squares model for a genome-wide quantitative trait association analysis in families. Hum Hered 71(1): 67-82.

Josse J, Holmes S. 2014. Measures of dependence between random vectors and tests of independence. Literature review. Ithaca, NY: Cornell University Library. Available at `http://arxiv.org/pdf/1307.7383v3.pdf` (accessed November 22th, 2014).

Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, Lee JC, Schumm LP, Sharma Y, Anderson CA and others. 2012. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature 491(7422): 119-124.

Jiang Y, Conneely KN, Epstein MP. 2014. Flexible and robust methods for rare variant testing of quantitative traits in trios and nuclear families. Genet Epidemiol 38: 542-551.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M. 2016. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Research 44: D457-D462 .

Kent JT. 1982. Robust properties of likelihood ratio tests. Biometrika 69: 19-27.

Kenny EE, Pe'er I, Karban A, Ozelius L, Mitchell AA, et al. 2012. A genome-wide scan of Ashkenazi Jewish Crohn's disease suggests novel susceptibility loci. PLoS Genet 8(3): e1002559.

Kim J., Zhang Y, Pan, W. 2016. Powerful and Adaptive Testing for Multi-trait and Multi-SNP Associations with GWAS and Sequencing Data. Genetics 203: 715-731.

Kim J, Bai Y, Pan W. 2015. An adaptive association test for multiple phenotypes with GWAS summary statistics. Genet Epidemiol 39: 651-663.

Klei L, Luca D, Devlin B, Roeder K. 2008. Pleiotropy and principal components of heritability combine to increase power for association analysis. Genet Epidemiol 32: 9-19.

Korte A, Vilhja lmsson BJ, Segura V, Platt A, Long Q, Nordborg M. 2012. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. Nat Genet 44: 1066-1071.

Kim S, Morris NJ, Won S, Elston RC. 2010. Single-marker and two-marker association tests for unphased caseacontrol genotype data, with a power comparison. Genet Epidemiol 34(1): 67-77.

Kwee LC, Liu D, Lin X, Ghosh D, Epstein MP. 2008. A powerful and flexible multilocus association test for quantitative traits. Am J Hum Genet 82: 386-397.

Kwak I, Pan W. 2016. Adaptive gene- and pathway-trait association testing with gwas summary statistics. Bioinformatics 32: 1178-1184.

Lee S, Emond MJ, Bamshad MJ, Barnes KC, et al. 2012. Optimal Unified Approach for Rare-Variant Association Testing with Application to Small-Sample Case-Control Whole-Exome Sequencing Studies. Am J Hum Genet 91: 224-237.

Li B, Leal S M. 2008. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 83: 311-321.

Li J, Tseng GC. 2011. An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. Ann Appl Stat 5: 994-1019.

Londono D, Buyske S, Finch SJ, Sharma S, Wise CA, Gordon D. 2012. TDT-HET: a new transmission disequilibrium test that incorporates locus heterogeneity into the analysis of family-based association data. BMC Bioinform 13(1): 13.

Liang K, Zeger S. 1986. Longitudinal data analysis using generalized linear models. Biometrika 73: 13-22.

Lin DY, Tang ZZ. 2011. A general framework for detecting disease associations with rare variants in sequencing studies. Am J Hum Genet 89: 354-367.

Lin J, Zhu H, Knickmeyer R, Styner M, Gilmore J, Ibrahim JG. 2012. Projection regression models for multivariate imaging phenotype. Genet Epidemiol 36: 631-641.

Lu Q, Hu Y, Sun J, Cheng Y, Cheung KH, Zhao H. 2015. A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. Scientific Reports, 5, Article number: 10576.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. 2009. Finding the missing herita- bility of complex diseases. Nature 461: 747-753.

Mantel N. (1967). The detection of disease clustering and a generalized regression approach. Cancer Research 27: 209-220.

Maity A, Sullivan PF, Tzeng J. 2012. Multivariate phenotype association analysis by marker-set kernel machine regression. Genet Epidemiol 36: 686-695.

Mahajan A, Go MJ, Zhang W, Below JE, Gaulton KJ, Ferreira T, Horikoshi M, Johnson AD, Ng MCY, Prokopenko I and others. (2014). Genome-wide trans-ancestry

meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. Nat Genet 46(3): 234-244.

Minas C, Curry E, Montana G. 2013. A distance-based test of association between paired heterogeneous genomic data. Bioinformatics: btt450.

Moerkerke B, Vansteelandt S, Lange C. 2010. A doubly robust test for gene-environment interaction in family-based studies of affected offspring. Biostatistics 11:213225. Pan W. 2009. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. Genet Epidemiol 33: 497-507.

Mühleisen TW, Leber M, Schulze TG, Strohmaier J, Degenhardt F, Treutlein J, Mattheisen M, Forstner AJ, Schumacher J, Breuer R and others. 2014. Genome-wide association study reveals two new risk loci for bipolar disorder. Nat Commun 5: 3339.

Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ 2010. Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. PLoS Genet 6: e1000888.

Nielsen DM, Ehm MG, Weir BS. 1998. Detecting marker-disease association by testing for HardyWeinberg disequilibrium at a marker locus. Am J Hum Genet 63(5): 1531-1540.

Pan W. 2009. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. Genet Epidemiol 33: 497-507.

Pan W. 2011. Relationship between genomic distance-based regression and kernel machine regression for multi-marker association testing. Genet Epidemiol 35: 211-216.

Pan W, Kim J, Zhang Y, Shen X, Wei P. 2014. A powerful and adaptive association test for rare variants. Genetics 197: 1081-1095.

Pan W, Kwak IY, Wei P. 2015. A Powerful Pathway-Based Adaptive Test for Genetic Association with Common or Rare Variants. American Journal of Human Genetics 97: 86-98.

Park Y, Sarkar A, Bhutani K, Kellis M. 2017. Multi-tissue polygenic models for transcriptome-wide association studies. bioRxiv. doi: http://dx.doi.org/10.1101/107623.

Pasaniuc B, Zaitlen N, Shi H, Bhatia G, Gusev A, Pickrell J, Hirschhorn J, Strachan DP, Patterson N and Price,AL. 2014. Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. Bioinformatics 30: 2906-2914.

Plagnol V, Howson JM, Smyth DJ, Walker N, et al. 2011. Genome-wide association analysis of autoantibody positivity in type 1 diabetes cases. PLoS Genet 7: e1002216.

Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, Boehnke M, Abecasis GR, Willer CJ. 2010. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics 26(18): 2336-2337.

Purcell SM, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ and others. 2007. Plink: a toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet 81: 559-575.

Qian M, Shao Y. 2013. A likelihood ratio test for genomeawide association under genetic heterogeneity. Ann Hum Genet 77(2): 174-182.

Roeder K, Bacanu SA, Wasserman L, Devlin B. 2006. Using linkage genome scans to improve power of association in genome scans. Am J Hum Genet 78: 243-252.

Rotnitzky A, Jewell NP. 1990. Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. Biometrika 77: 485-497.

Schaid DJ. 2010a. Genomic similarity and kernel methods I: advancements by building on mathematical and statistical foundations. Hum Hered 70: 109-131.

Schaid DJ 2010b. Genomic similarity and kernel methods II: methods for genomic information. Hum Hered 70: 132-140.

Schizophrenia Working Group of the Psychiatric Genomics Consortium. 2014. Biological insights from 108 schizophrenia-associated genetic loci. Nature 511(7510): 421-427.

Schifano ED, Epstein MP, Bielak LF, Jhun MA, Kardia SL, Peyser PA, Lin X. 2012. SNP set association analysis for familial data. Genet Epidemiol 36: 797-810.

Sejdinovic D, Sriperumbudur B, Gretton A, Fukumizu K. 2013. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. Annals of Statistics 41: 2263-2291.

Székely GJ., Rizzo ML, Bakirov NK. 2007. Measuring and testing dependence by correlation of distances. Annals of Statistics 35: 2769-2794.

Shen L, Kim S, Risacher SL, Nho K, Swaminathan S, West JD, Foroud T, Pankratz N, Moore JH, Sloan CD and others. 2010. Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: a study of the ADNI cohort. Neuroimage 53: 1051-1063.

Spearman, CE. 1904a. The proof and measurement of association between two things. Am J Psychology 15: 72-101.

Sun J, Zheng Y, Hsu L. 2013. A unified mixed-effects model for rare-variant association in sequencing studies. Genet Epidemiol 37: 334-344. Therneau TM. 2012. Mixed effects Cox models. R-package description. URL: `http://cran.r-project.org/web/packages/coxme/vignettes/coxme.pdf`.

Teslovich TM, et al. 2010. Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466: 707-713.

Torres JM, Barbeira AN, Bonazzola R, Morris AP, Shah KP, Wheeler HE, BellG, Cox NJ, Im HK. 2017. Integrative cross tissue analysis of gene expression identifies novel type 2 diabetes genes. bioRxiv. doi: https://doi.org/10.1101/108134

Therneau TM, Grambsch PM, Pankratz VS. 2003. Penalized survival models and frailty. J Comput Graph Stat 12: 156-175.

Tzeng JY, Zhang D, Pongpanich M, Smith C, McCarthy MI, Sale MM, Worrall BB, Hsu FC, Thomas DC, Sullivan PF. 2011. Studying gene and gene-environment effects of uncommon and common variants on continuous traits: a marker-set approach using gene-trait similarity regression. Am J Hum Genet 89: 277-288.

Wang X, Lee S, Zhu X, Redline S, Lin X. 2013. GEE-based SNP set association test for continuous and discrete traits in family-based association studies. Genet Epidemiol 37: 778-786.

Wang K. 2016. Boosting the Power of the Sequence Kernel Association Test by Properly Estimating Its Null Distribution. Am J Hum Genet 99: 104-114.

Wang, Y, Liu, A, Mills, JL, Boehnke, M, Wilson, AF., Bailey-Wilson, JE., Xiong, M, Wu, CO, Fan, R. 2015. Pleiotropy analysis of quantitative traits at gene level by multivariate functional linear models. Genet Epidemiol 39: 259-275.

Wang Z, Xu K, Zhang X, Wu X, Wang Z. 2017. Longitudinal SNP-set association analysis of quantitative phenotypes. Genet Epidemiol 41: 81-93.

Wellek S, Ziegler A. 2012. Cochran-Armitage test versus logistic regression in the analysis of genetic association studies. Hum Hered 73(1): 14-17.

Wessel J, Schork NJ. 2006. Generalized genomic distance-based regression methodology for multilocus association analysis. Am J Hum Genet 79: 792-806.

Wright F, Sullivan P, Brooks A, Zou F, Sun W, Xia K, Madar V, Jansen R, Chung W, Zhou YH, et al. 2014. Heritability and genomics of gene expression in peripheral blood. Nat Genet 46: 430-437.

Wigginton JE, Cutler DJ, Abecasis GR. 2005. A note on exact tests of Hardy-Weinberg equilibrium. Am J Hum Genet 76: 887-893.

Wittke-Thompson JK, Pluzhnikov A, Cox NJ. 2005. Rational inferences about departures from Hardy-Weinberg equilibrium. Am J Hum Genet 76: 967-986.

Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X. 2010. Powerful SNP-set analysis for case-control genome-wide association studies. Am J Hum Genet 86: 929-942.

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 89: 82-93.

Xiong M, Zhao J, Boerwinkle E. 2002. Generalized T2 test for genome association studies. Am J Hum Genet 70: 1257-1268.

Xu G, Lin L, Wei P, Pan W. 2016. An adaptive two-sample test for high-dimensional means. Biometrika 103: 609-624.

Xu Z, Shen X, Pan W; Alzheimers Disease Neuroimaging Initiative. 2014. Longitudinal analysis is more powerful than cross-sectional analysis in detecting genetic association with neuroimaging phenotypes. PLoS One 9: e102312.

Xu Z, Pan W. 2015. Approximate scorebased testing with application to multivariate trait association analysis[J]. Genet Epidemiol 39(6): 469-479.

Xu Z, Pan W. 2016. Binomial Mixture Model Based Association Testing to Account for Genetic Heterogeneity for GWAS. Genet Epidemiol 40(3): 202-209.

Yang Q, Wang Y. 2012. Methods for analyzing multivariate phenotypes in genetic association studies. J Probab Stat 2012: 652569.

Yang Q, Wu H, Guo CY, Fox CS. 2010. Analyze multivariate phenotypes in genetic association studies by combining univariate association tests. Genet Epidemiol 34: 444-454.

Zhou H, Pan W. 2009. Binomial mixture model-based association tests under genetic heterogeneity. Ann Hum Genet 73(6): 614-630.

Zhou X, Carbonetto P, Stephens M. 2013. Polygenic modeling with bayesian sparse linear mixed models. PLoS Genetics 9: e1003264.

Zhu X, Feng T, Tayo BO, Liang J, et al. 2015 Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. Am J Hum Genet 96: 21-36.

Zhu Z, Zhang F, Hu H, et al. 2016 Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nature Genet 48: 481-487.

Zou H., Hastie T. 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Series B 67: 301-320.